THE UNIVERSITY of York
CENTRE FOR HEALTH ECONOMICS
DEPARTMENT OF ECONOMICS AND RELATED STUDIES
NHS CENTRE FOR REVIEWS AND DISSEMINATION
YORK HEALTH ECONOMICS CONSORTIUM

A Formula for Distributing NHS Revenues Based on Small Area Use of Hospital Beds

Roy A. Carr-Hill
Geoffrey Hardman
Stephen Martin
Stuart Peacock
Trevor A. Sheldon
Peter Smith

A formula for distributing NHS revenues based on small area use of hospital beds

Results of a study commissioned from the University of York by the National Health Service Executive (formerly the National Health Service Management Executive)

Roy A. Carr-Hill
Geoffrey Hardman
Stephen Martin
Stuart Peacock
Trevor A. Sheldon
Peter Smith

September 1994

The Authors

All the authors are members of the University of York.
Roy Carr-Hill is a Senior Research Fellow in the Centre for Health Economics.
Geoffrey Hardman is a Research Fellow in the Centre for Health Economics.
Stephen Martin is a Research Fellow in the Institute for Research in the Social Sciences.
Stuart Peacock is a Research Fellow in the York Health Economics Consortium.
Trevor Sheldon is Director of the NHS Centre for Reviews and Dissemination.
Peter Smith is a Senior Lecturer in the Department of Economics and Related Studies.

Acknowledgements

A full list of acknowledgements is given in the introduction.

Address for Correspondence

Peter Smith
Department of Economics and Related Studies
University of York
York YO1 5DD

Further Copies

Further copies of this document (at price £12.50 to cover the cost of publication, postage and packing) from:

The Publication Secretary
Centre for Health Economics
University of York
York YO1 5DD

Please make cheques payable to the University of York. Details of other papers can be obtained from the same address, or telephone York (01904) 433648.

A formula for distributing NHS revenues based on small area use of hospital beds

PREFACE

This document reports the results of a study to examine the determinants of NHS hospital inpatient utilization commissioned by the National Health Service Executive. The purpose was to develop a formula which was suitable for distributing annually about £18 billion of Hospital and Community Health Service funds to health authorities in England. The report sets out the background to the study, describes the data on which it was based, explains the statistical methodology used, and presents the findings. The implications for revenue allocations to individual health authorities are not discussed in this report. A summary of contents and findings is given in Chapter 1.

CONTENTS

1.	**INTRODUCTION AND SUMMARY**		**1**
2.	**RESOURCE ALLOCATION IN THE NHS**		**8**
	2.1	A brief history of resource allocation in the NHS	8
	2.2	Criticisms of previous resource allocation methods	13
	2.3	Summary	18
3.	**THE BACKGROUND TO THIS STUDY**		**22**
	3.1	The principles of resource allocation	22
	3.2	A model of demand for health care	26
	3.3	Data sources	33
4.	**MEASUREMENT ISSUES**		**38**
	4.1	Measuring the need for health care	38
		4.1.1 Demography	39
		4.1.2 Health status	40
		4.1.3 Socio-economic conditions	44
	4.2	Measuring the supply of health care	47
		4.2.1 Accessibility of NHS inpatient services	48
		4.2.2 Accessibility of GP services	51
		4.2.3 Provision of nursing and residential homes	53
		4.2.4 Accessibility of private hospitals	54
		4.2.5 Higher level supply variables	55
	4.3	Measuring utilization of NHS resources	57
		4.3.1 The Hospital Episode Statistics	57
		4.3.2 Calculating utilization measures	63
	4.4	Conclusion	71
5.	**MODELLING NHS INPATIENT UTILIZATION**		**73**
	5.1	Estimating a small area model of utilization	73
	5.2	Developing a resource allocation formula	79
	5.3	Multilevel modelling of utilization	82
6.	**RESULTS**		**85**
	6.1	An acute specialties model	85
	6.2	The non-acute specialties models	93
	6.3	Sensitivity analysis	99
	6.4	Conclusions	102
7.	**CONCLUSIONS**		**106**
Appendix A		The creation of synthetic small areas	112
Appendix B		Census variables used in the study	114
Appendix C		Measuring accessibility	118
Appendix D		The statistical modelling strategy	122
Appendix E		Multilevel estimation	134
Appendix F		Options for further work	137

LIST OF ABBREVIATIONS

DHA	District Health Authority
DHSS	Department of Health and Social Security
DoH	Department of Health
ED	Enumeration District
FHSA	Family Health Service Authority
GMS	General Medical Services
GP	General Practitioner
HCHS	Hospital and Community Health Services
HES	Hospital Episode Statistics
NHS	National Health Service
ICD	International Classification of Diseases
LBW	Low Birth Weight
ML	Multilevel
OLS	Ordinary Least Squares
OPCS	Office of Population Censuses and Surveys
RAWP	Resource Allocation Working Party
RHA	Regional Health Authority
SIR	Standardized Illness Ratio
SMR	Standardized Mortality Ratio
SSA	Synthetic Small Area
SSR	Standardized (permanent) Sickness Ratio
UPA	Underprivileged Area (score)
2SLS	Two stage least squares

1. INTRODUCTION AND SUMMARY

1.1 This document describes the findings of a statistical analysis of the determinants of utilization of National Health Service inpatient facilities by small areas in England. It forms part of a study by the University of York commissioned by the NHS Executive. The primary purpose of the study was to "improve the sensitivity of the current formula for allocating Hospital and Community Health Service funds to Regional Health Authorities".

1.2 In addition, the study was to pursue four secondary objectives, namely:

a) to address the possibility of developing a national formula for general practitioner fundholding procedures;

b) to assist the four Thames Regional Health Authorities produce an acceptable sub-regional Pan-Thames formula;

c) to investigate the relationship between age and sex and the use of hospital beds;

d) to determine some of the substitution and complementary relationships between different types of health care.

This report describes the study's findings on the main objective, and also covers secondary objectives (c) and (d). The development of a formula for general practitioner (GP) fundholding procedures and of a formula for the Thames Regions are the subject of separate reports.

1.3 Throughout the study, the York team was guided by two advisory groups appointed by the Department of Health: the Technical Group advised on technical matters, and the Steering Group advised on policy matters. All important decisions relating to measurement, modelling and interpretation of results were

made in consultation with members of the Steering and Technical Groups. The study team is grateful for their advice and support.

1.4 The report starts with a brief history of resource allocation in the NHS (Section 2.1). Since the Resource Allocation Working Party (RAWP) report in 1976, a large part of the Hospital and Community Services budget has been distributed to Regional Health Authorities on the basis of formulae. However, although intuitively attractive, the RAWP formula was not directly based on empirical evidence, and in 1990 it was superseded by a revised formula based on the "Review of RAWP", a statistical analysis of hospital utilization. The Review was criticised on a number of grounds (Section 2.2), and one of the principal objectives of this study is to subject utilization data to a more rigorous statistical analysis.

1.5 Chapter 3 sets out the background to this study, and the model of demand for health care on which it is based. The basis of the study was an analysis of the utilization that small areas make of NHS inpatient facilities. The 4,985 small areas are "synthetic wards", which are aggregations of electoral wards, with average population of about 10,000. Fundamental to the analysis was the idea that both population needs and supply of health care facilities play an important part in shaping demand for health care. A simple mathematical model of demand for health care is set out, and it is noted that, because needs influence both utilization and supply, the usual statistical regression methods may be inappropriate. Section 3.3 then describes the data that were available to the study team to make the model of demand operational.

1.6 Chapter 4 discusses the problem of measuring the concepts of need, supply and utilization. Needs cannot be measured directly. However, numerous potential indicators of needs were available. These can be considered under three headings: demography, health status and broader socio-economic conditions. Demography

was tackled throughout the study by standardizing for age and sex all variables which were thought to depend on demographic considerations. Health status was measured in four ways: mortality, limiting long-standing illness, permanent sickness and low birth weight. A large number of variables were extracted from the 1991 Census of Population to reflect socio-economic conditions.

1.7 Central to the measurement of supply of health care is the notion of accessibility. Deriving a measure of accessibility entails reconciling the attractiveness of facilities, their distance from the population of interest, and the influence of competition from neighbouring populations. This was achieved using the methods of spatial interaction modelling (Appendix C). In all, four dimensions of supply were modelled: accessibility of NHS hospitals, accessibility of GP services, provision of nursing and residential homes, and accessibility of private hospitals.

1.8 In the first instance, measures of utilization were derived from the 1990/91 Hospital Episode Statistics (HES) database (Section 4.3). These were subsequently augmented with data from the 1991/92 HES. Each episode was assigned to the synthetic ward of residence of the patient, and so the total utilization of a ward could be inferred in terms of numbers of episodes and numbers of bed days. In addition, using an analysis of specialty costs prepared by East Cheshire Statistical Analysis Consultancy, it was possible to infer a measure of the costs of each episode. This could be used as a measure of intensity of utilization. The analysis was undertaken for acute and non-acute specialties separately.

1.9 Chapter 5 describes the statistical methods used. Having noted that ordinary regression methods are inappropriate, an alternative approach, based on the method of "two stage least squares" is set out. This entails adjusting the supply variables to take account of the fact that they may be influenced by needs before entering them into the regression equation. Section 5.1 describes the method, and explains how the large number of potential needs indicators available to the study

team was narrowed down to a manageable number. Throughout, there was a need to ensure that the model was well specified, in the sense that it did not breach any of the usual assumptions of statistical analysis.

1.10 Having identified satisfactory models of utilization, there was a requirement to infer a formula for resource allocation. At this stage, the study team and its advisors recognised that legitimate health care needs can affect utilization in two ways: first, directly, and second indirectly, to the extent that current supply already reflects legitimate health care needs. In order to model both direct and indirect effects, it was necessary to conduct regressions of utilization on variables proven to reflect health care needs alone. The rationale for this approach - which represents a conceptual advance in this area - is set out in Section 5.2.

1.11 The intention in this study was to detect the national average response to needs, after adjusting for variations in supply not justified by needs. However, the analysis might be confused by systematic policy differences between administrative areas, such as District Health Authorities (DHAs) or Family Health Service Authorities (FHSAs). That is, utilization patterns might tend to cluster within administrative areas. If this is the case, usual statistical methods might be invalid. The study therefore also used multilevel modelling techniques to adjust for systematic inter-District effects (Section 5.3), and it is the coefficients from this latter exercise which are preferred.

1.12 The principal results of the study are set out in Chapter 6 The model identified for acute specialties (including births) incorporates the following variables:

- all causes standardized mortality ratio (ages 0-74);
- proportion of those of pensionable age living alone;
- proportion of dependants living in households with only one carer;
- standardized limiting long-standing illness ratio (ages 0-74);
- proportion of economically active persons unemployed.

1.13 A variety of models were developed for the non-acute specialties (geriatrics and mental illness). The mental illness model includes:

- proportion in households headed by a lone parent;
- proportion of dependants with no carers;
- proportion in households with head born in New Commonwealth;
- proportion of those of pensionable age living alone;
- all causes standardized mortality ratio for ages 0-74;
- proportion of adult population permanently sick.

It proved impossible to develop a model for mental handicap using the study methodology.

1.14 Numerous alternative specifications of these models were explored, and an extensive sensitivity analysis was undertaken, testing the robustness of the models by:

(a) examining variations between Regions;
(b) examining variations between specialty groups;
(c) restricting the analysis to elective admissions;
(d) restricting the analysis to over-65 age groups;
(e) examining variations between "high" and "low" needs areas;
(f) exploring alternative measures of the costs of episodes.

1.15 Chapter 7 sets out our conclusions and recommendations. In summary, they are as follows:

(a) that the models identified can be used as the basis for formulae for allocating inpatient Hospital and Community Health Service (HCHS) funds to Regions;

(b) that consideration should be given as to how to treat mental handicap,

community and outpatient services in any national formula;

(c) that the formulae could be used as a basis for setting District targets;

(d) that, where necessary, a system for adjusting allocations to Districts in the light of local circumstances should be retained;

(e) that consideration should be given to initiating a national cohort study as an alternative means of measuring the link between social circumstances and health care needs in individuals;

(f) that the study dataset should be released to other researchers as soon as possible.

1.16 The study team wishes to note its grateful thanks to numerous people, without whose help the study could not have been completed. As noted above, the help of the Steering Group and Technical Group was essential to the conduct of the study, and is gratefully acknowledged. Keith Derbyshire, Peter Dick and Nick York were our contacts in the NHS Executive, and they offered unfailing and generous support. Also from the Department of Health, Peter Goldblatt and Frank O'Hara gave us much helpful advice and practical support, as did John Charlton and Dave Foote at the Office of Population Censuses and Surveys. The cost estimates central to the study were produced by Ken Johnson, whose untimely death occurred just as the project was starting.

1.17 Completion of the study would have been impossible without invaluable consultancy support from the following:

Michael Borowitz	*Battelle Medical Research and Policy Centre*
Harvey Goldstein	*Institute of Education*

Chris Orme	*University of York*
John Rasbash	*Institute of Education*
Annabelle Shakespeare	*Battelle Medical Research and Policy Centre*
Min Yang	*Institute of Education*

We also benefited from the help of numerous others, including George Davey Smith of the University of Glasgow, Di Jackson of Leicestershire Health, Professor Brian Jarman of St Mary's Hospital Medical School, Gwyn Bevan of London Economics, Nick Mays of the University of Ulster, and some anonymous referees. At York, we had invaluable secretarial support from Kerry Atkinson and advice from Professors Alan Maynard and Alan Williams. The heroic efforts of John Byrne and Dave Atkin in the Computing Service were also essential for the success of the project.

2. RESOURCE ALLOCATION IN THE NHS

2.1 A major problem confronting most government programmes is how to distribute their spending between geographical areas. The National Health Service is no exception. For the first 30 years of its life, allocations to geographical areas were predominantly incremental: that is, based on the previous year's allocation plus an element for growth. Funds for general practice were also largely distributed on this incremental basis until the mid 1960s, since when a capitation element has predominated. However, since 1976 NHS expenditure on hospital and community services has been distributed according to formulae designed to secure some notion of geographical equity.

2.2 In 1992/93 the NHS spent a total of £27.9 billion in England (Department of Health, 1993). Of this, about £5.2 billion (19%) were spent on General Practitioner services; £21.4 billion (77%) on Hospital and Community Health Services; and £1.0 billion (4%) on central health services and central administration. The subject of this report is the distribution of expenditure on Hospital and Community Health Services (HCHS) in England. It therefore excludes consideration of the distribution of General Practitioner funds. This Chapter reviews the history of attempts to allocate HCHS revenue expenditure in England, and describes the strengths and weaknesses of the chosen methods.

2.1 A brief history of resource allocation in the NHS

2.3 Towards the end of the 1960s, government ministers had become acutely aware of the inequities and inefficiencies that arose when revenues were distributed incrementally. The high concentration of teaching hospitals in London gave cause for particular concern. As a result, a Resource Allocation Working Party (RAWP) was set up by the Secretary of State to recommend a system for the allocation of resources which was responsive to the health needs of the population, and to identify and correct inequalities in the existing pattern of resource distribution. It

considered how NHS capital and revenue resources should be allocated to the Regional Health Authorities (RHAs), by them to the now defunct Area Health Authorities, and subsequently to the District Health Authorities (DHAs).

2.4 The Working Party reported in 1976 (Department of Health and Social Security, 1976). It recommended distributing revenue resources on the basis of population, weighted according to two fundamental criteria: first, adjustments were to be made for perceived differences in the *need* for health care; and second, account was taken of the unavoidable geographical differences in *costs* of providing services. The principle of a *weighted capitation* formula has remained intact since the RAWP report, and informs the current report.

2.5 In developing a measure of relative need, the first step was to acknowledge the role played by demographic characteristics. Accordingly, the population of each area was disaggregated by age and sex, thereby permitting adjustment for the considerable variations in use of NHS resources made by different age/sex groups.

2.6 It was recognized, however, that there were social and economic determinants of need in addition to the purely demographic. In principle, these should be measured by some index of morbidity. However, the Working Party recognized the difficulty of developing such an index. In particular, they rejected the solution of using resource utilization as a proxy for health care need because of the supposed strong influence of service provision on utilization. Consequently, they recommended using the Standardized Mortality Ratio (SMR) as an index of morbidity, and therefore as a proxy for need. The SMR is defined as the number of *observed* deaths in an area as a percentage of the *expected* deaths in the area, where expected deaths are calculated by applying national age/sex specific death rates to the age/sex structure of the local population. SMRs can be calculated for any subgroup of the population or specific conditions, although random variations due to small numbers often make such disaggregation meaningless at a small area

level.

2.7 A decision was made to break down health care into broad categories of conditions, and the index of relative need for care for each category was determined by applying the all ages condition-specific SMR to the population of an area. Thus the SMR was taken to indicate relative need for services relating to a particular condition for every age group.

2.8 Finally, the condition-specific indices were combined by multiplying the SMR-weighted number of persons in each age/sex group by the national average number of bed days used by a person in that age/sex group diagnosed as suffering from that condition. Bed days were therefore a "currency" whereby need in different conditions could be combined into a single index. This process generated a notional total use of bed days by the population in an area, assuming utilization conformed to the national average, after adjusting for local need, as indicated by the SMRs. Algebraically, the equation can be represented as follows:

$$RA_i = \sum_j SMR_{ij} \left(\sum_k BEDS_{jk} POP_{ik} \right) \qquad \textbf{(2.1)}$$

where RA_i is the allocation to area i; SMR_{ij} is the SMR of condition j in area i; $BEDS_{jk}$ is the national number of bed days required by age/sex group k diagnosed with condition j; and POP_{ik} is the population in area i in age/sex group k.

2.9 The RAWP formula was accepted by the government as a basis for distributing HCHS revenue funds to Regional Health Authorities. After taking account of teaching responsibilities, cross boundary flows and higher costs of service delivery in the south east, it was used to generate revenue "targets" for each of the RHAs, which were to be secured by a progressive shift of funding away from RHAs above target towards those below target over a number of years. It was found that the four Thames Regions and Oxford were considerably above target, so a process

of redistributing resources from the south east towards the north began (Allsop, 1984). The process was designed to occur slowly so that the redistribution did not cause too much disruption. Despite the initial uproar from the losers, the scheme gained a degree of acceptance in that it was open and seen to be generally fair. Adjustments were subsequently made to various factors, but the basic principles of RAWP were left intact until 1988. In addition, most Regions used the principles of RAWP to distribute revenue funds within their Region.

2.10 In 1985 the NHS Management Board was asked to review the operation of the RAWP formula. In December 1986 they produced an Interim Report which outlined some general principles and a programme of further work (Department of Health and Social Security, 1986). The final report was issued in the summer of 1988 (Department of Health and Social Security, 1988).

2.11 In announcing the review, the Secretary of State made it clear that

> "the underlying principle of RAWP, that of securing equal opportunity of access to health care for people in equal need, is not in question ... However, as Regions move closer to their RAWP targets, it becomes increasingly important that the targets themselves should reflect relative need as fairly as possible. The review will therefore ... look at the scope for improving the measurement of need." (DHSS, 1986, Annex C)

The concern of the Interim Review Report was that there should be

> "An analysis of the proxies for need for health services, including different forms of SMRs, social and other factors carried out on a small area basis." (DHSS, 1986, para 3.15, p.4)

2.12 The stated aim of the review was therefore to improve the accuracy with which the formula measured relative need, because the Regions were gradually converging towards their targets and it was thought that fine tuning was required. The formula was to be evaluated and - if necessary - changed according to two basic principles stated in both the interim and final reports: first, that no change should

be made to the formula unless clearly justified; and second, that the formula should remain as stable, robust and as simple and straightforward as perceived "fairness" permitted for national purposes (Interim Report, paras 2.13 and 2.15). These are valuable principles and should be borne in mind in assessing the results of this study.

2.13 The majority of the work carried out for the Review by Coopers and Lybrand (1988) was based on ordinary least squares regression analyses of the determinants of hospital utilization in small areas (electoral wards). A variety of functional forms were explored, including additive and multiplicative models. The explanatory variables tested included a measure of service availability (occupied beds weighted by distance to the electoral wards), and a variety of SMRs and deprivation factors. In the event, the Review recommended several changes to the formula, of which the most important were as follows:

(1) The population weighting by age/sex was to be changed: an 85+ age band was to be included and sex stratification dropped.

(2) The SMR weighting factor was to be changed, with the abandonment of condition-specific SMRs. The new age-weighted population was to be multiplied by the all-condition SMR for people aged below 75 (denoted SMR75). Moreover, the elasticity attached to the SMR was reduced from 1 to 0.44. The move away from the one-to-one correspondence between SMR and need presumed by RAWP had the effect of reducing the importance of variations in SMR between the regions in the formula by as much as 2% of the final target (Mays, 1989).

(3) A social deprivation factor was to be included in the needs weighting, in the form of the Underprivileged Areas (UPA8) index. The aim was to weight the population additionally by the degree of social deprivation

experienced in the region, which is supposed to be another predictor of need and hence of resource utilization (Jarman, 1983, 1984).

The recommended equation gave a needs index for area i ($NEEDS_i$) as follows:

$$NEEDS_i = k.SMR75_i^{0.44}.e^{0.0026*UPA8_i} \qquad (2.2)$$

This index replaced the simple SMR needs index used in RAWP. Regional allocations were then found by multiplying by age/sex specific utilization rates, as in RAWP.

2.14 The government did not accept all of the Review's findings, and eventually implemented a modified version of the formula which dropped the indicator of social deprivation, and altered the coefficient on the under-75 SMR to 0.5. At the same time, changes were made to allocations for teaching responsibilities and the adjustment for higher costs in London. At the subregional level, some Regions have used the Review's age cost curves and the all causes under 75 SMR as an index of need, while others have commissioned and implemented entirely new methods, with the aim of reflecting the particular circumstances and characteristics of the local population (for example, Balarajan, 1990).

2.2 Criticisms of previous resource allocation methods

2.15 Although the RAWP formula gained a considerable amount of acceptance, the large redirection of funds it implied inevitably generated tensions which needed addressing Some of these tensions were pragmatic, in the sense that the increasing demands on NHS budgets led to health authorities everywhere coming under fiscal pressure and questioning their allocations. However, there were also shortcomings in principle in the RAWP approach, of which not all were satisfactorily addressed by the Review (Mays and Bevan, 1987). This Section reviews the principal criticisms of the resource allocation methods used to date.

2.16 The RAWP formula was based on the following syllogism: morbidity measures need; mortality is a proxy for morbidity; therefore mortality is an indicator for need. Several studies have challenged the second statement: that mortality is a good proxy for morbidity (for example, Forster 1976; 1978). In particular, there have been demonstrations of a statistical association between various spatial measures of morbidity and socio-economic factors over and above any relationship with mortality. Indeed, it is now commonly agreed that RAWP's choice of mortality as a surrogate for morbidity was incomplete - although it is of course unclear what should be put in its place (Carr-Hill, Maynard and Slack, 1990). The use of all-ages SMRs was also criticised on the grounds that it was dominated by the large number of deaths in older age groups. Mortality in the 75+ age group may simply be the inevitable consequence of ageing, and may not reflect the level of morbidity (and therefore the need for health care) within the population as a whole.

2.17 Also implicit in the RAWP formula was the assumption of a direct one-to-one relationship between resources and need, as indicated by mortality rates. It has been argued that this assumption had no empirical justification, and may be inappropriate (Barr and Logan, 1977; Butts 1986).

2.18 A further crucial step in the RAWP formula was the assumption that HCHS revenues could be allocated on the basis of national use of hospital beds by patients suffering from conditions associated with a particular SMR category. This assumption was vulnerable to four fundamental criticisms. First, the link between ICD codes (the basis for condition-specific SMRs) and patient conditions is not straightforward. Second, the use of bed days is in any case only a rough proxy for the total consumption of resources in hospital and community. Ideally, some measure of the total costs associated with the use of NHS resources is required. Third, the formula is intrinsically conservative because it assumes that the existing national allocation of resources between conditions is satisfactory. In particular, it

takes no account of suppressed demand for NHS resources. And fourth, the formula assumed that the existing national allocation of resources between demographic groups is satisfactory.

2.19 The Review of RAWP sought to address concern about the measurement of need. SMRs were still used as the fundamental index of need. However, disaggregation by condition was abandoned in favour of an age-limited all-cause SMR. In addition, the results of the statistical analysis indicated that there was not a one-to-one relationship between SMR and need, as assumed by RAWP. Moreover, the proposed augmentation of SMRs with an index of social deprivation sought to capture aspects of need not reflected in mortality measures.

2.20 Much of the criticism of the Review of RAWP focused on methodological shortcomings. One fundamental problem is that summary measures of need in an area do not necessarily reflect the sum of the needs of individuals living in the area (the ecological fallacy). It is usual to argue that the use of small areas as the unit of analysis is justified because wards are likely to be reasonably homogeneous in terms of population characteristics, and that in any case there is no better unit of analysis available (Hume and Womersley, 1985; Leavey and Wood, 1985; Scott-Samuel, 1984; Townsend, Phillimore and Beattie, 1987). Carr-Hill (1990) argues that there is considerable variation *within* wards (almost as much as *between* wards) so that analysis should ideally be carried out at a lower level, such as enumeration districts.

2.21 However, the use of small areas at any level does not overcome the difficulty that it may be inappropriate to infer causal relationships at the individual level on the basis of associations found at the group level. For, whilst the end point of analysis is a formula to allocate resources to large geographical areas, those resources are intended to lead to the provision of the most appropriate patient care for individuals. The derivation of an appropriate formula should therefore in

principle be based on the link between individual need and use. As Carstairs and Morris (1989) point out, such purism is currently impracticable, requiring access to cohort studies tracing the relationship between individual characteristics and individual resource use. Nevertheless, although not undertaken at present in the UK, it should be noted that such studies are used to inform health care resource allocation in other countries, such as France (Charraud and Mormiche, 1986).

2.22 The Review used the number of inpatient episodes as a measure of utilization, and so did not seek to model non-inpatient care. The episode measure of utilization makes no allowances for variations in case mix and severity, and therefore in resource use, between different areas. It should be noted that the Review found that models based on bed days had little explanatory power. Furthermore, no attempt was made to disaggregate into different specialties, even though the determinants of need might be quite different in different specialties.

2.23 The Review team recognized that, in moving to bed utilization as a measure of need, it needed to adjust for the impact of supply of NHS resources on use. However, historical supply is itself a function of need (not least because previous allocation formulae are likely to have reflected need). As a result, it may be inappropriate to employ the ordinary least squares (OLS) regression techniques used in the Review. Supply and demand are almost certainly jointly determined by a complex feedback process, and so more advanced statistical methods must be used (Mays, 1989; Sheldon and Carr-Hill, 1992). Moreover, the Review only considered supply of NHS hospital services. In practice, utilization of HCHS resources is likely to be influenced by the existence of other services, such as primary care, private health care and personal social service provision.

2.24 More generally, utilization of hospital resources is influenced by patient and professional behaviour which is not necessarily related to need (Morgan, Mays and Holland, 1987). In practice, therefore, as well as influencing need, social

circumstances are also likely to influence the predisposition to seek health care, and the nature of that care. Thus, socio-economic considerations enter the model twice; first as a factor in determining morbidity and therefore need; and second as a determinant of the care sought, independent of any clinical need. Disentangling these two roles of social characteristics is almost certainly beyond the scope of simple statistical models, given current availability of data.

2.25 The Review did not address the issue of whether existing patterns of health care are optimal, and its results were vulnerable to the influence of the practices of service providers. Thus, for example, areas with efficient health care services - perhaps with an emphasis on care outside hospital - could be deemed to require smaller resources than equally needy areas with less efficient services (Milner and Nichol, 1988). Equally, the review did not address the issue of whether existing patterns of use were appropriate: that is, whether some demographic and socio-economic groups over- or under-use services relative to other groups with the same health needs.

2.26 The Review's statistical methods were criticised on a number of grounds (Sheldon, 1990; Sheldon and Carr-Hill, 1992). The team did not appear to subject their models to standard tests for statistical misspecification, such as tests for endogeneity of variables; tests of functional form; and tests of model stability - although this was done later by some of the team's members (Hancock, Holden and Swales, 1991). The analysis was moreover confined to a subset of observations, which may have been biased. More generally, Sheldon, Davey Smith and Bevan (1993) note that the search for an empirically based capitation formula will always be limited by the range and reliability of the available data.

2.3 Summary

2.27 The NHS now has extensive experience of using formulae to distribute health care resources between geographical areas. The principles of weighted capitation proposed by the RAWP have been widely accepted. However, the following problems still require attention:

(1) identification of suitable indicators of need for health care;

(2) taking full account of the impact of social circumstances on the need for health care;

(3) taking into account the impact of supply of services when inferring the determinants of utilization, and distinguishing between variations in supply which reflect need and those which are determined by non-need factors, such as historical regional differences;

(4) identification of good indices of the costs of health care;

(5) taking account of possible differences in policies and health care priorities between geographical areas;

(6) taking account of possible differences in efficiency between geographical areas;

(7) assessing whether the use of existing national allocations of resources to care groups is appropriate as the basis for a formula;

(8) ensuring that technically adequate statistical methods are used for estimation purposes.

2.28 In order to address these issues, better theory, data and methodology than hitherto are needed. This study sought to address all three issues within the severe time constraints imposed.

Theory The approach taken by the York team included explicit consideration of the supply-utilization link and the multidimensional impact of need. Our modelling drew on a synthesis of the best available knowledge in this area. Moreover, the team was able to involve in the study some of the critics of previous work. The theoretical framework is set out in Chapter 3.

Data We were able to exploit a more comprehensive hospital inpatient dataset, incorporating national coverage, division between specialty groups, and a variety of measures of resource use. In addition, we has available the 1991 Census, which is contemporaneous with the inpatient data, and which included a question on Limiting Long Standing Illness for the first time in a Census. The available data are described in Chapter 4.

Methodology The study team was able to call upon a wide range of statistical expertise, including extensive experience of analysing complex systems of interdependent relationships, of handling different levels of aggregation amongst variables, and of testing that statistical models are well specified. The methodology adopted is described in Chapter 5.

It is moreover planned that this study's dataset should be made available to the Health Service and academic communities for further analysis.

2.29 Of course there remain significant limitations which will be pointed out at the appropriate stages in the development of the models.

References

Allsop J (1984) *Health policy and the National Health Service*, Longman, London.

Balarajan R (1990) *Social deprivation and age adjustment ratios for South West Thames Region*, Epidemiology and Public Health Research Unit, University of Surrey, Guildford.

Barr A and Logan R F L (1977) Policy alternatives for resource allocation, *The Lancet*, 1, 994-996.

Butts M (1986) Questioning basic assumptions, *Health Service Journal*, 19 June, 826-827.

Carr-Hill R A (1990) RAWP is Dead - Long Live RAWP, *Health Policy 13,* 135-44.

Carr-Hill R A, Maynard A K and Slack R (1990) Morbidity Variation and RAWP *Journal of Epidemiology and Community Health* 44(4), 271-273.

Carstairs V and Morris R (1989) Deprivation and Mortality: an alternative to social class? *Community Medicine*, 11(3), 210-219.

Charraud A and Mormiche P (1986) *Disparités de consummation medicales enquete sanité 1980-1981*. Les collections de l'INSEE, serie M118, 1986.

Coopers and Lybrand (1988) *Integrated Analysis for the Review of RAWP,* Coopers and Lybrand, London.

Department of Health (1993) *Departmental Report*, HMSO, London.

Department of Health and Social Security (1976) *Sharing Resources for Health in England*, Report of the Resource Allocation Working Party, HMSO, London.

Department of Health and Social Security (1986) *Review of the Resource Allocation Working Party Formula: interim report by the NHS Management Board*, DHSS, London.

Department of Health and Social Security (1988) *Review of the Resource Allocation Working Party Formula: final report by the NHS Management Board*, DHSS, London.

Forster D P (1976) Social class differences in sickness and general practitioner consultation, *Health Trends*, 8, 29.

Forster D P (1978) Mortality as an indicator of morbidity in resource allocation, in Brotherston J (ed) (1978) *Morbidity and its relationship to resource allocation*, Welsh Office, Cardiff.

Hancock K E, Holden D R and Swales JK (1991) Consistency of the spatial allocation of NHS resources: an econometric analysis, *Applied Economics*, 23, 1623-1636.

Hume S M and Womersley J (1985) Analysis of death rates in the population aged 60 years and over of Greater Glasgow by postcode sector of residence", *Journal of Epidemiology and Community Health*, 39(4), 357-365.

Jarman B (1983) Identification of under-privileged areas *British Medical Journal,* 286, 705-709.

Jarman B (1984) Underprivileged areas: validation and distribution of scores *British Medical Journal*, 289, 1587-1592.

Leavey R and Wood J (1985) Does the underprivileged area index work? *British Medical Journal*, 291, 709-711.

Mays N (1989) NHS resource allocation after the 1989 White Paper: a critique of the research for the RAWP review, *Community Medicine*, 11, 3, 173-186.

Mays N and Bevan G (1987) *Resource Allocation in the Health Service*, Occasional Paper in Social Administration 81, Bedford Square Press, London.

Milner P and Nichol J (1988) Revising RAWP, *The Lancet* 2, 1195.

Morgan M, Mays N and Holland W (1987) Can Hospital Use be a Measure of Need for Health Care? *Journal of Epidemiology and Community Health* 41(4), 269-274.

Scott-Samuel A (1984) The need for primary health care: an objective indicator, *British Medical Journal,* 288, 457-458.

Sheldon T A (1990) The Problems of Using Multiple Regression for Modelling the Demand, Supply and Use of Health Services *Journal Public Health Medicine* 12(3/4) 213-215.

Sheldon T A and Carr-Hill R A (1992) Resource allocation by regression in the NHS: a statistical critique of the RAWP review, *Journal of the Royal Statistical Society (Series A)*, 155(3), 403-420.

Sheldon T A, Davey Smith G and Bevan G (1993) Weighting in the dark: resource allocation in the new NHS, *British Medical Journal* 306, 835-839.

Townsend P, Phillimore P and Beattie A (1987) *Health and Deprivation: Inequality and the North*, Croom Helm.

3. THE BACKGROUND TO THIS STUDY

3.1 The principal objective of this study was to suggest a formula suitable for distributing HCHS revenue resources in an equitable manner to health authorities. The means to that end is a statistical analysis of the utilization of hospital services by small areas. This Chapter sets the scene for that study. Section 3.1 examines the principles on which a resource allocation formula might be based. Section 3.2 sets out the conceptual model of supply of and demand for health care resources on which this study is based. Finally, Section 3.3 describes the data sources available to the study team.

3.1 The principles of resource allocation

3.2 The founding principles of the National Health Service included the notion of equality of access for those with equal need. This principle was acknowledged by the RAWP team, which interpreted its terms of reference as being:

> "to reduce progressively, and as far as feasible, the disparities between the different parts of the country in terms of the opportunity for access to health care of people at equal risk" (DHSS, 1976).

The principle of equality of access for those in equal need has been reaffirmed both in the 1988 review and in the guidelines for this project.

3.3 Thus, some concept of equity underlies any attempt to allocate resources rationally. However, the analysis of various notions of equity is notoriously complex (Pereira, 1993). For example, in commenting upon the SHARE formula, the Scottish analogue of the RAWP formula, Mooney (1982) outlined seven possible interpretations of equity:

(1) equality of expenditure per capita;

(2) equality of inputs (resources) per capita;

(3) equality of input for equal need;

(4) equality of (opportunity of) access for equal need;

(5) equality of utilization for equal need;

(6) equality of marginal met need;

(7) equality of health.

3.4 There has been general agreement that Mooney's first two formulations would constitute an inadequate response to inequity. Clearly, unless one believes that all people and groups should be treated equally regardless of inter-personal variation, some other factor has to be used to guide the equity criterion. At the other end of the spectrum, the prospect of attaining equality in health is seen as utopian, and most commentators would agree that the choice is between criteria (3), (4), (5) and (6).

3.5 The RAWP report explicitly argued - and most commentators have agreed - that one should not aim to secure equity criterion (5) - equal utilization rates for a presumed equal need - because there may be many legitimate means of meeting health care needs. The report also noted the practical problem of taking account of historical supply variation when using utilization as a criterion for assessing equity. Much of the subsequent debate about resource allocation has, of course, been about how to adjust for the influence of supply variations when using utilization data.

3.6 People do not in general demand health care because they want treatment, but because they expect an improvement in health status. Some health economists have therefore argued that an equitable distribution of resources should equalize in terms of marginal met need, in line with Mooney's criterion (6). That is, the aim should be the economist's notion of allocative efficiency, securing the maximum output possible given the resources available (Culyer, 1988). The problem with this argument is the practical difficulty of measuring and comparing the improvements in health status brought about by clinical intervention (Carr-Hill and Sheldon, 1992).

3.7 The quote at the start of this Section suggests that the RAWP team had in mind Mooney's fourth equity formulation. However, it is impossible to guarantee equality of access to those in equal need through the medium of a resource allocation formula alone. As a result, the recommended methodology could only implement the third criterion of equity: equality of input for equal need. Clearly, only if all DHAs have the same set of health care priorities and offer the same levels of efficiency can this be translated into criterion (4).

3.8 In addition to these theoretical considerations, Le Grand (1987) has argued that a practical interpretation of equity in health care must take account of the processes by which health states are determined, in particular the extent to which they arise from factors beyond individual control. A simple version of this thesis is already incorporated in the current formulae by the age/sex adjustment. It is assumed that the considerably heavier use by older age groups is brought about by their greater need. Le Grand proposes an extension based on the argument that disparities in health which reflect variations in the circumstances of different population groups are inequitable. This implies that one of the aims of resource allocation should be to give health authorities the ability to correct for the effects of unavoidable social circumstances. However, the problem of disentangling avoidable from unavoidable determinants of health care needs is beyond the scope of this study.

3.9 In translating these principles of equity into action, it is necessary to develop a measure of the need for health care. The difficulties this gives rise to are discussed in more detail in Section 4.1. To date, SMRs have been used as indices of health care needs, and, as Mays and Bevan (1987) report, it seems that they encapsulate many of the social determinants of ill health. However, there is evidence that social considerations over and above those reflected in mortality and morbidity may nevertheless have an important bearing on the demand for health care. This study accepts the need to test for the impact of a range of socio-economic considerations in the determination of demand for health care.

3.10 The above discussion raises the issue of where to draw the limits to any consideration of social welfare provision. Although this study is concerned with the Hospital and Community Health Services, the demand for such services may be heavily influenced by provision in parallel public services, such as primary medical care, local authority personal social services and public sector housing. Whether a resource allocation formula for the HCHS should adjust for provision in these other welfare services is a fundamental matter of principle. Similarly, the populations of different areas might make different use of private medical care. To what extent should any resource allocation formula take account of the impact this might have on the use of NHS resources?

3.11 The study team recognizes the complexity of these issues. Ideally one might envisage a unified resource allocation model for health care services provided by all agencies, in particular integrating allocations to HCHS and General Practice. When data and time permit, it may become possible to address this important agenda. However, the immediate need was to develop an equitable and manageable formula with which to allocate HCHS resources alone. In doing so we attempted to take account of some aspects of parallel provision. For example, the provision of many of the complementary welfare services, and the use of private health care, are likely to be reflected - albeit in some complex way - in the socio-economic characteristics of an area, and these form an intrinsic component of our model (see Section 4.1).

3.12 Moreover, in analysing demand for health care, we sought to consider explicitly the impact of provision of GP services, residential homes and private health care on hospital utilization, as explained in Section 4.2. The model we develop therefore describes the impact of health and social needs on resource utilization after adjusting for supply considerations. We subsequently sought to isolate that variation in supply which can be accounted for by variations in health care needs. However, we are inclined to the view that the incorporation of the parallel

provision of other welfare services *directly into a HCHS formula* is in general infeasible without further work, and we have therefore omitted explicit measures of non-HCHS supply from our recommended formulae.

3.13 We believe that previous resource allocation methods have been informed by principles which are in line with society's requirements of the NHS. The principle of equity they embody is that areas in equal "need" should receive equal resource allocations. Underlying the methods is the notion that the benchmark for need is the existing use of resources at the national level by particular age/sex population groups (Sheldon, Davey Smith and Bevan, 1993). Local allocations are therefore adjusted for local population structure. These are then in turn adjusted by a local health care needs index.

3.14 Although this procedure is essentially conservative, in that for example it accepts existing allocations of resources between age/sex groups at the national level, it has gained general acceptance, partly because of a reluctance to open up the issue of the share of the NHS cake between age groups. Most criticism has focused on the actual implementation of the underlying ideas. The most discussed problem with RAWP was the appealing but unsubstantiated claim that SMRs reflected need for services in a one-to-one relationship. The main criticisms of the Review of RAWP centre on the methodology used to establish the empirical link between needs variables and the use of NHS resources. As a result, this study does not challenge the principles underlying previous work. Instead, it seeks to develop a formula based on sound evidence and more robust statistical methodology.

3.2 A model of demand for health care

3.15 In order to proceed, it is necessary to make explicit the assumed process underlying the demand for health care. Clearly any representation of this complex phenomenon is likely to be highly schematic. However, we offer the following

model of the demand for health care, illustrated in Figure 3.1, in the belief that it captures most of the salient features relevant to the required analysis of utilization. The underlying socio-economic and demographic characteristics of the population give rise to health care needs, in terms of morbidity. Through some imperfectly understood process, these needs give rise to a certain level of demand for health care. The determinants of this demand are however not just the health characteristics of the population. Social characteristics may also influence demand independently of health status. Thus, social needs and the expectations of the population may have influences on demand over and above any narrow consideration of health needs.

3.16 In the same way, the availability of health and related welfare services can affect expectations, and therefore the demand for health care. More specifically, in this study we take the view that it is likely that the *perceived* availability of local services will be a major determinant of local demand. For example, when a general practitioner decides whether or not to refer a patient to a consultant, she may be influenced by the time the patient will have to wait for treatment. And what matters in the referral process is the GP's *perception* of the expected waiting time, and not necessarily any objective measure of waiting time. This emphasis on perceived accessibility - which is, of course, unobservable - is important because it draws attention to the potential for very local variations in effective supply, even though the physical supply of beds available to GPs may be a constant across a district. Supply can influence utilization in other ways. For example, the level of supply will influence the extent to which demand for health care can be met in situations where there is excess demand for health care. In addition, there is a body of research which suggests that "supplier-induced demand" might be an important consideration (Cromwell and Mitchell, 1986).

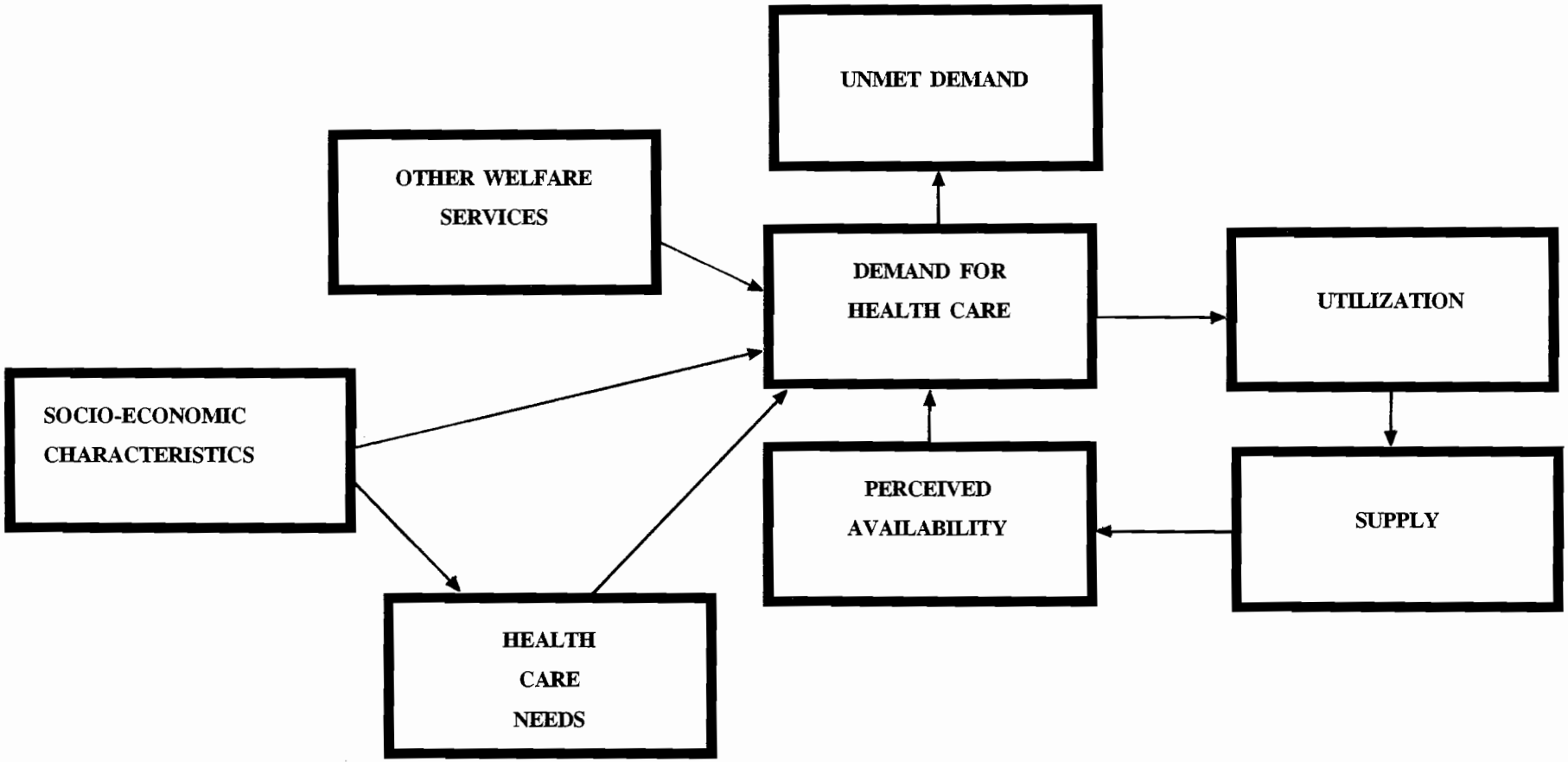


Figure 3.1: A model of demand for health care

3.17 In short, an underlying need for health care in the population is augmented by social circumstances and expectations to generate demand for health care. In the light of this demand and the local political process, NHS services are provided. The adequacy of the supply response will then affect future expectations, and therefore future demand. Moreover, the level of utilization over time affects the physical availability of services, which in turn affects perceived availability, and so the process continues. It is also possible that the activities of the NHS might have an effect on the underlying health status of the community. Thus there is a feedback from supply to demand. The actual use of NHS facilities is therefore a dynamic process, with many of the links in Figure 3.1 containing time lags.

3.18 Modelling such a dynamic process is clearly complex, and calibrating a satisfactory empirical model impossible, given the current limited availability of data. In particular, there is a lack of adequate time series data, and many of the existing measures of utilization and supply are very crude. However, the situation described above can be represented algebraically in a simplified way as follows. The level of utilization U_i in locality i is a function of health care needs N_i, the perceived availability of local services P_i, the actual level of services provided S_i and socio-economic and demographic considerations X_i:

$$U_i = f_1(N_i, P_i, S_i, X_i) \qquad (3.1)$$

The level of perceived availability is probably itself a function of actual availability and other considerations X:

$$P_i = f_2(S_i, X_i) \qquad (3.2)$$

and health care needs are a function of socio-economic and demographic conditions

$$N_i = f_3(X_i) \qquad (3.3)$$

Finally, supply is a function of utilization (including both historical patterns of use

and current demand) and historical needs, via:

$$S_i = f_4(U_i, U_i^{-1}, N_i^{-1}, X_i) \qquad (3.4)$$

where the (-1) superscript refers to levels of utilization or needs at some time in the past.

3.19 This system of equations and, in particular equations 3.1 and 3.4 suggest that supply and utilization are jointly determined. That is, they are both affected by the same processes, either at the same time or with lags. In the terminology of econometrics, the variables are *endogenous*. The extent to which supply variables are correctly treated as endogenous has been discussed in the literature (Cromwell and Mitchell, 1986; Sheldon and Carr-Hill , 1992). Much depends on the level of analysis being adopted. Thus, given historical administrative arrangements in England, the supply of beds made available at the *district* level may indeed have been influenced by historical aggregate needs within districts. However, it is not necessarily the case that physical supply of beds is endogenous at the *ward* level. Nevertheless, what affects utilization is perceived supply, and this will vary according to the characteristics of the local area (electoral ward) and the GPs acting as agents within it. Therefore there is every reason to suppose that endogeneity exists even at the small area level.

3.20 Note that whilst the needs variables included in equation 3.1 represents current needs and those in equation 3.4 represent past needs, this does not affect the argument about simultaneity between supply and utilization which is the crux of the problem in estimating unbiased coefficients for the relation between needs and utilization. However, in practice, the time series data implied by (3.4) are not available. Therefore, we are forced to assume that current needs are a proxy for past needs, and that current utilization is a proxy for past utilization. Then, abandoning the distinction between perceived and actual supply, we obtain the following equations:

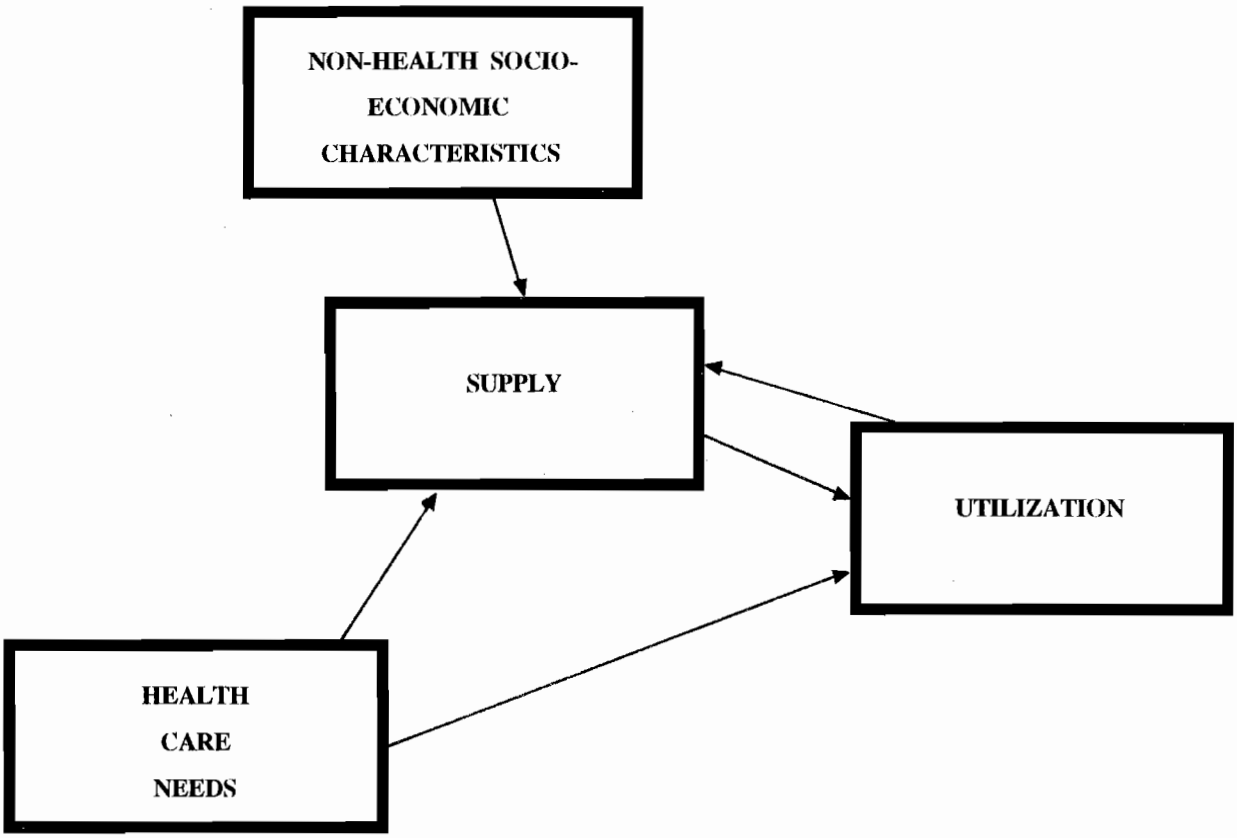


Figure 3.2: The simplified model of demand for health care

$$U_i = g_1(N_i, S_i) \tag{3.5}$$

and

$$S_i = g_2(N_i, U_i, X_i) \tag{3.6}$$

The simplified model implied by (3.5) and (3.6) suggests that utilization is a function of health care needs and supply. Supply in turn is determined by needs, utilization and other socio-economic characteristics not directly related to health care needs. The model is illustrated in Figure 3.2.

3.21 Because of the assumptions we have had to make, this model is almost certainly a simplification of reality, in the sense that it fails to capture the subtle interaction of needs, supply and utilization over time. Yet even if the model specified in equations (3.5) and (3.6) were theoretically sound, the narrow availability of data and limitations of statistical methods might constrain the ability to build a meaningful empirical model of the demand for health care. In particular, given that the relationship (3.6) exists, it is almost certainly inappropriate to seek to estimate the utilization equation (3.5) by means of ordinary least squares regression, as was attempted in the Review of RAWP, because the simultaneous determination of U and S would lead to biased estimates of regression coefficients. As a result, it is necessary to examine the possibility of endogeneity, and if it exists to use more advanced statistical estimation techniques, such as two-stage least squares. This insight is central to the current study, and is explained in more detail in Section 5.1.

3.22 As explained in the next Section, the units of analysis will be small geographical areas. Although this approach offers the best practical method of examining the relationship between needs and utilization using routinely collected data, it suffers from a further complicating factor. It might reasonably be assumed that medical and administrative policies (as adopted, say, by hospitals, DHAs or FHSAs) have an influence over a geographical area that is wider than the small area unit of analysis. This is particularly likely in Districts outside the metropolitan areas, which are primarily supplied by one hospital. In these circumstances, it is quite plausible to suggest that there are systematic "District" effects which influence utilization across a number of observational units.

3.23 This poses a problem for statistical estimation because the fundamental assumption of randomly distributed error terms may be violated. Therefore, as a supplementary analytic tool, this study employs multi-level modelling techniques to test for higher level supply effects (Paterson and Goldstein, 1991). In effect,

multilevel modelling adds a further set of supply variables into the model: the policy effects of administrative areas. The multilevel analysis enables us to explore the robustness of the models developed using regression techniques, and can be used to investigate the extent and causes of inter-authority variation, as explained in Section 5.1.

3.3 Data sources

3.24 In order to make the preceding principles operational it is necessary to derive measures of need, supply and utilization. These are discussed in Chapter 4. The measures must then be incorporated into a mathematical model which can be estimated using statistical techniques, the subject of Chapter 5. As a prelude, this Section gives an outline of the various data sources that were available to the study team, the nature of which influenced the modelling methodology adopted.

3.25 The fundamental unit of analysis was a "synthetic ward", the smallest area of analysis available to the study team. This level of analysis was chosen by the study's Technical Group in advance of commissioning the study, as explained in Appendix A. In summary, the Technical Group judged that some electoral wards were too small to offer reliable estimates of variables used in the study. Wards with small populations (less than 5,000) were therefore amalgamated with other neighbouring wards until a unit with a population of at least 5,000 was derived. Details of the aggregation procedure are given in Appendix A.

3.26 In deciding to adopt a small area unit of analysis, an important consideration for the Technical Group was the "ecological fallacy": the identification of relationships which are statistically significant at one area level of aggregation (say a health authority) but which do not reflect any underlying relationships obtaining at a different level of aggregation (perhaps the individual). The use of relatively homogeneous small area units of analysis is intended to minimize the biases

attributable to this phenomenon. However, it should be noted that populations of 5,000 can be extremely heterogeneous, and that heterogeneity may still be a problem even if it were possible to carry out the analysis at the level of the enumeration district. It can be argued that the problems caused by the ecological fallacy can only be incontrovertibly eliminated when the analysis is undertaken at the level of the individual (Carr-Hill, 1990).

3.27 The aggregation process resulted in the formation of 4,985 synthetic wards covering the whole of England, with average population 9,643. Each synthetic ward lies within a local authority district, and therefore a Family Health Service Authority (FHSA). Furthermore, it was found that most synthetic wards lie within a single District Health Authority (DHA). Where this was not the case, the synthetic ward was allocated to the DHA covering the largest proportion of its population. Only 92 of the synthetic wards required such assignment.

3.28 The data associated with each synthetic ward were derived from six principal sources:

(1) OPCS data on population, mortality and low weight births, provided at the synthetic ward level;

(2) Data from the 1991 Census of Population, available at the local authority ward level;

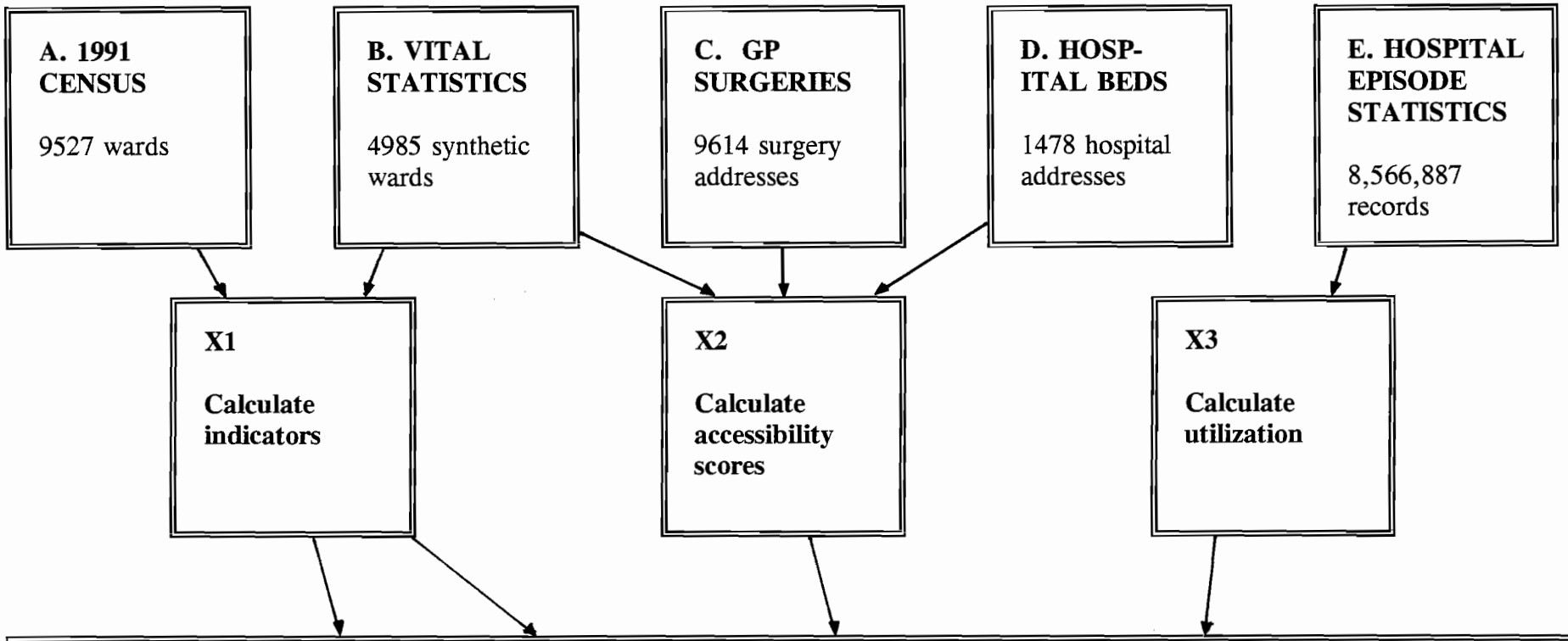
(3) A database of all hospitals with over 19 beds in England provided by the Department of Health Management Executive (1,478 records);

(4) A database of all General Practitioner main surgeries in England provided by the Department of Health Management Executive (9,671 records, of which 57 were not usable because of inadequate addresses);

(5) The 1990/91 Hospital Episode Statistics (HES), comprising 8,566,887 records of all completed inpatient and day case hospital episodes (as explained later, the 1991/92 HES were made available towards the end of the study, after the bulk of the model development had been undertaken);

(6) The 1990/91 Health Service Indicators package, giving summary indicators related to health service resources, activities and performance in DHAs and FHSAs.

3.29 The contents of these data sources were processed in a variety of ways to describe the characteristics of individual synthetic wards. The transformation processes are described in detail in Chapter 4. The links between the various data sources and the master database are illustrated in Figure 3.3. The first level shows the six principal data sources. The second level represents the various procedures required to associate the raw data with synthetic wards. And the third level shows the outline structure of the database of synthetic wards. Section M1 contains demographic data from the mid-1991 population estimates and administrative data; Section M2 contains data relating to socio-economic conditions in the synthetic ward; Section M3 contains health-related data; Section M4 contains ward-specific data relating to the supply of NHS and private health care facilities; Section M5 contains utilization information derived from the HES data; and Section M6 contains data associated with the administrative areas (DHAs and FHSAs) in which the synthetic ward lies. This database forms the central information resource for the study.



MASTER DATASET (4985 records)					
M1. Housekeeping data	M2. Socio-economic data	M3. Health data	M4. Supply data	M5. Utilization data	M6. Multi-level data
Ward code DHA code FHSA code Population Grid reference	Socio-economic indices	Mortality Illness Low birth weight	Hospital access GP access Private hospitals Nursing homes	Episodes Bed days Estimated costs Standard costs	Expenditure Supply Policies

Figure 3.3: Representation of the construction of the master dataset

References

Carr-Hill R A (1990) RAWP is dead: long live RAWP, in Culyer A J, Maynard A and Posnett J (1990) *Competition in health care: reforming the NHS*, Macmillan Press, Basingstoke.

Carr-Hill R A (1992) *Methods of revenue allocation for Wales*, Welsh Office, Cardiff.

Carr-Hill R A and Sheldon T A (1992) Rationality and the use of formulae in the allocation of resources to health care, *Journal of Public Health Medicine*, 14, 117-126.

Cromwell J and Mitchell J B (1986) Physician-induced demand for surgery, *Journal of Health Economics*, 5, 293-313.

Culyer A J (1988) "Are inequalities inevitable" in *Acceptable Inequalities,* IEA Health Unit.

Department of Health and Social Security (1976) *Sharing Resources for Health in England*, Report of the Resource Allocation Working Party, HMSO, London.

Le Grand J (1987) "Equity, health and health care" in *Three Essays on Equity*, Welfare State Programme Discussion Paper No. 23, London School of Economics

Mays N and Bevan G (1987) *Resource Allocation in the Health Service*, Occasional Paper in Social Administration 81, Bedford Square Press, London.

Mooney G (1982), *Equity in Health Care: Confronting the Confusion.* Discussion paper No 11/82, Health Economics Research Unit, Aberdeen.

Paterson L and Goldstein H (1991) New statistical methods for analysing social structures: an introduction to multilevel models, *British Educational Research Journal*, 17(4), 387-393.

Pereira J (1993) What does equity in health mean?, *Journal of Social Policy*, 22, 19-48.

Sheldon T A and Carr-Hill R A (1992) Resource allocation by regression in the NHS: a statistical critique of the RAWP Review, *Journal of the Royal Statistical Society, Series A,* 155, 403-420.

Sheldon T A, Davey Smith G and Bevan G (1993) Weighting in the dark: resource allocation in the new NHS, *British Medical Journal* 306, 835-839.

4. MEASUREMENT ISSUES

4.1 In order to build useful empirical models of the demand for health care, it is necessary to seek to measure numerous phenomena relating to both supply and demand. This Chapter considers measurement issues under three headings. Section 4.1 considers potential determinants of the need for health care; Section 4.2 examines measures of the supply of health care and related services; and Section 4.3 describes the measures of utilization used in this study. Some selective descriptive statistics are given. However, it should be borne in mind that the number of variables considered was very large, so it is infeasible in this report to offer more than an outline of the statistical properties of most of the variables. The full dataset used in the study is available for scrutiny.

4.1 Measuring the need for health care

4.2 As discussed in Chapter 3, the determinants of the need for health care are complex and poorly understood. Indeed, the assumption underlying this study is that need *per se* is unmeasurable, and can only be inferred by examining the link between health status and socio-economic circumstances and utilization. It is moreover usually impossible to determine whether socio-economic characteristics affect demand indirectly (through their influence on health status) or directly (through their influence on demand for any given level of health). And some direct influences - such as the presence of a carer at the patient's home - might be considered legitimate reasons for variations in utilization, while others - such as raised expectations of the NHS amongst certain social classes - might be considered as biasing provision towards those classes. These are weighty issues which cannot be addressed within the framework of this study. However, they should be borne in mind when examining the results.

4.3 In practice, we wish to identify measurable socio-economic indicators which appear to influence the legitimate use of NHS hospital services, on the assumption

that these indicators reflect unmeasurable underlying need. There is a considerable amount of existing research in this area. In summary, three types of population characteristics appear to influence the demand for health care: demography, morbidity and material socio-economic conditions. These are now considered in turn.

4.1.1 Demography

4.4 Levels of morbidity, and therefore of the need for health care, clearly vary with an individual's age and sex. As a result, other things being equal, areas with different age/sex profiles will in general have different health care needs. For each ward, this study had available the mid-1991 population estimates prepared by the Office of Population, Censuses and Surveys (OPCS) for males and females in 18 five-year age groups. From a modelling perspective, these highly disaggregated data give rise to the problem of capturing age and sex characteristics of an area in a manageable number of variables, as any simple variable of demographic structure (for example, percentage aged over 74) is likely to be crude and arbitrary. Wherever possible, therefore, all variables that vary substantially with age and sex were standardized. For most indices, 18 age and two sex groups were used.

4.5 Standardization can take two forms: direct and indirect. Direct standardization entails applying *local* rates (of illness, mortality, utilization, etc.) to *national* population profiles. Algebraically, this process is as follows:

$$DSR_i = \sum_{j=1}^{18} \sum_{k=1}^{2} m_{ijk} P_{jk} \qquad (4.1)$$

where DSR_i is the directly standardized index in ward i, j refers to age group, k refers to sex, m_{ijk} is the age/sex specific rate in ward i, and P_{jk} is the national proportion of population in age/sex group {jk}.

4.6 Using indirect standardization, the *national* rates M_{jk} are applied to *local* population sizes p_{ijk}. This yields

$$EN_i = \sum_{j=1}^{18} \sum_{k=1}^{2} M_{jk} p_{ijk} \qquad (4.2)$$

where EN is the expected number of observations of the phenomenon of interest in the area. The indirectly standardized rate is then the ratio of *observed* number of observations to this expected number.

4.7 In practice, for all indices studied for which both types of age/sex standardization could be undertaken, the difference between direct and indirect standardization was found to be negligible. We therefore used indirect standardization throughout in order to be consistent with standardized mortality ratios, which are indirectly standardized. This yielded a ratio of observed to expected number of observations in a ward, given its demographic characteristics. Standardization is intended to obviate the need to include any age or sex variables in our model of demand. In practice, however, standardization might not fully remove the effect of demography on demand, so - once a preferred model was identified - a large number of demographic variables were added to the chosen model to check that no residual impact of age or sex remained unexplained.

4.1.2 Health status

4.8 Four types of variable directly related to health status were available at the ward level:

(1) Standardized mortality ratios (SMRs) for all causes of death, as provided by OPCS;

(2) Standardized illness ratios (SIRs), the standardized proportion of persons reporting long term limiting illness in the 1991 Census;

(3) Standardized permanent sickness ratios (SSRs) - the standardized proportion of the population of working age permanently sick - derived from the Census;

(4) Low birth weight - the proportion of live and still births for which weights were recorded which were less than 2.5kg.

4.9 The SMR has long been established as an important variable in explaining variations in health care resource use. It reflects the cumulative morbidity and social experience of an area and provides a more stable, unbiased and comprehensive measure of morbidity than most measures of utilization. By way of illustration, Table 4.1 shows the correlations of a variety of ward-based SMRs with some selected socio-economic variables. Note the high correlations of all variables with SMRs for ages up to 75, and the relatively low correlations for the 75+ age group. It has been well established that mortality rates are highly correlated with the incidence of chronic diseases which are known to justify health service intervention (Mays & Bevan, 1987). Furthermore, the dying are amongst the heaviest users of services. A US study has indicated that as much as 28% of Medicare expenditure is spent on those in the last year of life (Lubitz and Prihoda, 1982). In the UK, it is not possible to get comparative estimates for individuals, but 23% of non-psychiatric beds were occupied by people who died before discharge in England and Wales in 1984 (Sheldon, Davey Smith and Bevan, 1993). Hence variations in mortality not only indicate variations in morbidity, but also variations in the great need for services caring for those with conditions associated with a high number of fatalities.

Proportion	SMR 0-64	SMR 65-74	SMR 75+	SMR 0-74	SMR ALL
In households with no car	0.781	0.627	0.195	0.781	0.568
Children with non-earning lone parents	0.699	0.543	0.151	0.687	0.488
Of working age permanently sick	0.693	0.640	0.305	0.735	0.611
Of working age unemployed	0.741	0.604	0.217	0.745	0.564
With limiting long term illness	0.574	0.496	0.277	0.591	0.517

Table 4.1: Correlations of SMRs with socio-economic variables

4.10 The current NHS national resource allocation formula weights the population according to national average age-specific utilization rates. The above discussion suggests that an important contribution to these rates arises from the utilization of people who subsequently die. These will of course will be predominantly in the older age groups. Therefore, if there are a higher than average number of deaths in younger age groups in a District, its need for health care resources would be underestimated if crude age-specific utilization rates were used. It is therefore vital to include SMRs alongside the weightings for age-specific utilisation rates to correct for variations in life expectancy.

4.11 The OPCS provided us with SMRs for each ward for the three calendar years 1990 to 1992 for three age groups: under 65, 65 to 74, and 75 and over. It was therefore possible to construct six SMRs, as follows: SMR65 (under 65), SMR75 (under 75), SMRALL (all ages), SMR65+ (65 and over), SMR75+ (75 and over), and SMR65-74 (65 to 74). In interpreting these data, it should be borne in mind that, because of the lower life expectancy of males, the SMRs for younger

age groups will tend to be dominated by mortality patterns amongst men. Ideally, we should use finer age bands so that indices based on roughly equal numbers of deaths in males and females could be constructed.

4.12 The 1991 Census of Population included for the first time a question about long term limiting illness: "does the person have any long-term illness, health problem or handicap which limits his/her daily activities or the work he/she can do?". Respondents were asked to include problems which were due to old age. Clearly, the question is self-reported, and is vulnerable to variations in reporting practice. For example, O'Donnell and Propper (1991) identify differences in reporting between social groups in response to general questions about morbidity, such as this Census question. However, unpublished work suggests that positive response to the question is strongly associated with GP utilization, and its consideration in this study is essential. We therefore constructed a standardized index based on the same age/sex groups as the SMR, reflecting the ratio of observed to expected numbers suffering long term limiting illness in a ward. We call this the Standardized Illness Ratio (SIR), and have calculated it for the same age groups as reported above for the SMR. In this Chapter, we report results for the SIR for all residents in a ward (including those living in institutions). As explained in Chapter 6, we subsequently tested the implications of restricting the SIR to persons living in households.

4.13 In addition, the Census included a question about permanent sickness amongst those of working age. The question asked about the person's economic activity in the week preceding the Census, and one possible response was "unable to work because of long term sickness or disability". Responses to this question permitted creation of a single Standardized Sickness Ratio (SSR), in which the age categories used were restricted to those of working age.

4.14 The incidence of low weight births (LBW) may reflect morbidity amongst mothers

and more general social considerations, such as maternal smoking. Clearly the 2.5 kg limit is to some extent arbitrary, and a local rate may reflect genetic factors unrelated to health. In addition, the variable may be vulnerable to variations in the quality of recording practice in different areas. Nevertheless, the proportion of births less than 2.5 kg may yield important information about patterns of morbidity in the population which is not reflected in any of the other health status variables considered here.

4.15 Table 4.2 indicates the correlations between the health status variables. In order to indicate the effects of the standardization process, the crude unstandardized proportion of persons reporting limiting long term illness is also included. With the exception of low weight births, the variables are highly correlated, although the correlation between SMR 0-74 and SMR 75+ is relatively small. Note in particular the very high correlation (0.96) between the all ages SIR and the SSR.

4.1.3 Socio-economic conditions

4.16 It is widely agreed that socio-economic conditions - most importantly related to economic deprivation - are important determinants of the need for health care, over and above considerations of intrinsic health status (Mays and Bevan, 1987). We have already noted a putative relationship between the presence of carers at home and the use for health care resources. And, in a different sense, high earning households may have a greater predisposition to use private health care, and therefore may not require the same level of NHS resources as lower income households with the same health status.

4.17 In practice, the link between socio-economic conditions and the need for health care is poorly understood, and many of the data - such as income levels - needed to gain a better understanding of the link are not available. It is therefore necessary to consider the socio-economic variables used to explore demand for

	SMR 0-64	SMR 65-74	SMR 75+	SMR 0-74	SMR ALL	Long_t illness	SIR 0-64	SIR 65-74	SIR 75+	SIR 0-74	SIR ALL	SSR	LBW
SMR 0-64	1.00	0.65	0.30	0.91	0.71	0.57	0.77	0.65	0.45	0.77	0.75	0.74	0.44
SMR65-74		1.00	0.46	0.90	0.79	0.50	0.69	0.70	0.57	0.71	0.71	0.68	0.33
SMR 75+			1.00	0.42	0.85	0.28	0.30	0.35	0.51	0.32	0.35	0.31	0.11
SMR 0-74				1.00	0.83	0.59	0.81	0.75	0.56	0.82	0.81	0.78	0.43
SMR ALL					1.00	0.52	0.65	0.64	0.64	0.66	0.69	0.64	0.31
Long_t ill						1.00	0.79	0.67	0.56	0.78	0.79	0.78	0.27
SIR 0-64							1.00	0.86	0.62	0.99	0.98	0.97	0.45
SIR 65-74								1.00	0.75	0.91	0.92	0.84	0.36
SIR 75+									1.00	0.66	0.72	0.62	0.23
SIR 0-74										1.00	0.99	0.97	0.44
SIR ALL											1.00	0.96	0.43
SSR												1.00	0.39
LBW													1.00

Table 4.2: Correlations between health variables

Key:
- SMR — Standardized mortality ratio (for various age groups)
- Long_t ill — Unstandardized ratio of those with limiting long term illness to total population
- SIR — Standardized limiting long term illness ratio (for various age groups)
- SSR — Standardized permanent sickness ratio for those of working age
- LBW — Low weight births (<2.5kg) as percentage of all births

health care as proxies for the underlying social causes of need, such as poverty. With this in mind, researchers and policy makers have developed a number of indices which seek to provide a deprivation score for geographical areas. Examples include the Jarman and Townsend indices of deprivation. These indices have been criticised for their lack of theoretical basis, and the combination of highly correlated variables. We saw no need to use them in this study, as we were able to use all the components of a deprivation index separately, without constraining the way in which they were combined.

4.18 A large number of variables from the Census were created for use in this study. The full list is given in Appendix B. In summary, the variables cover the following aspects of social and economic circumstances:

Housing Tenure
Housing Amenities
Car ownership
Overcrowding
Ethnic origin
Elderly living alone
Lone parents
Students
Migrants
Unemployment
Educational qualifications
Social class
Concealed families
Non-earning households

4.19 Clearly it is possible to propose a large number of measures in addition to those given in Appendix B. However, we believe that, for the purposes of this study, the range of issues covered is likely to be sufficient to capture the important social causes of the need for health care. Indeed it is noteworthy that, because there is high correlation between many of the variables, the exclusion of a variable from the chosen models does not necessarily mean that the phenomena it is measuring are not captured in the models. It is important to be clear that the variables are

acting as proxies for unmeasurable health care needs characteristics of the ward. Thus, for example, lack of car ownership or unemployment may be reflecting aspects of poverty rather than simply measuring narrow economic status.

4.20 The Census is a valuable source of social information. However, it is undertaken only every 10 years, and so no variables derived from it can be annually updated. They may therefore be unsatisfactory for use in a formula for annual revenue allocations if relative social conditions alter during the period between Censuses. In addition, the 1991 Census is thought to have suffered from high levels of under-recording compared with previous Censuses. The under-recording was almost certainly not uniform between areas, and may therefore have lead to bias. It would therefore be desirable to incorporate socio-economic data from other sources. However, we were unable to find any source that was routinely available at the level of local authority ward. Most importantly, the regular unemployment series assembled by the Department of Employment will not be made available at 1991 ward level until 1995.

4.2 Measuring the supply of health care

4.21 To some extent the distinction between *social* determinants of demand and *health care supply* determinants of demand is artificial. For example, concealed amongst the social variables there may be an underlying predisposition to purchase private health insurance, which - if it could be isolated - might be more appropriately thought of as a health services supply variable. Similarly, the presence of a carer at a dependant's home may in some circumstances offer a substitute for inpatient care. In this Section we therefore confine the discussion to direct measures of health care resource provision. It should be noted, however, that modelling supply of health care resources at the ward level is severely constrained by the availability of data. Many of the supply characteristics within the NHS are best described at the DHA or FHSA level, rather than the ward level. As a result, as

the principal tool in the multilevel modelling work described in Section 5.2, we have constructed a number of variables related to DHA and FHSA policies and service provision. However, we have been able explicitly to model four features of the provision of health and related services which might be thought *prima facie* to have an influence on demand and which vary significantly between wards.

4.22 The four measures of the supply of health care services at the ward level are: the physical proximity of hospital beds; the provision of GPs; the provision of nursing home services; and the provision of private hospital beds. This Section discusses each in turn. It concludes with a summary of the DHA and FHSA level variables used in the multilevel analysis.

4.2.1 Measuring accessibility of NHS hospital inpatient services

4.23 There is a widespread belief, promulgated in the RAWP report, that the supply of hospital services affects demand for those services. A fundamental need in this study is therefore to develop a measure of the *perceived availability* of NHS inpatient services to a particular ward. This measure should incorporate three elements: the inherent *attractiveness* of services; their *proximity* to the population of interest; and the *competition* for use of the services from populations in other wards. The traditional method of treating such concepts is to develop a measure of the *accessibility* of the ward to NHS services. This is achieved here using the ideas of spatial interaction modelling described in detail in Appendix C.

4.24 As Appendix C explains, the accessibility A_i of zone i to hospital facilities is given by the rather complex expression

$$A_i = g\left(\sum_d T_{id}\right)/P_i = g\sum_d B_d\, S_d\, f(c_{id}) = g\sum_d \left(\frac{S_d\, f(c_{id})}{\sum_r P_r\, f(c_{rd})} \right) \qquad \textbf{(4.3)}$$

where T_{id} is the number of interactions (hospital episodes per year) between residential zone i and hospital d;

P_i is the population of zone i;

S_d is some measure of the size or attractiveness of hospital d;

B_d is a balancing factor for hospital d;

c_{id} is some measure of distance (or time) between i and d;

f(.) is a distance decay or deterrence function;

g is a constant.

4.25 Although superficially opaque, equation (4.3) - which is similar to the supply measure used in the Review of RAWP - can be interpreted simply as the ratio of population (weighted by distance) to hospital size (weighted by distance). If the measure of hospital size is taken to be beds, the equation is directly analogous to the familiar "beds per head" ratio, but takes account of distance and competition from other wards.

4.26 In order to obtain an accessibility score, it is necessary to find measures of the size of inpatient facilities, and their location. This study used a database of 1,478 hospitals in England (excluding Special Health Authorities) which gave the numbers of beds in five specialties (acute, maternity, geriatric, mental handicap and psychiatric). Clearly more refined measures of hospital size would be desirable, but it was infeasible to pursue these within the time constraints of this study. Choice of measures for P_i and S_d was therefore straightforward: total population and available beds serve as the only practical proxies.

4.27 It should be noted that the accessibility variable is only intended to capture the extent to which the actual physical supply of beds and of consultants are perceived as accessible to the population of different wards. The study methodology does not require that this measure should be adjusted for relative need in the population, as the socio-economic determinants of supply are explicitly considered at the

modelling stage. Indeed, the purpose of building the models described in Chapter 5 is to isolate those health care needs variables which can be considered legitimate determinants of supply. It is therefore unnecessary and illogical to seek to adjust supply measures for relative needs before the models have been estimated.

4.28 The database also gave the address of the hospital. From this, a grid reference could be inferred for each hospital, based on the grid reference of its postcode. The grid reference of the population-weighted centroid of each synthetic ward was also available. It was therefore possible to calculate an approximation to the straight line distance between every hospital and every ward. Of course, ideally the measure of distance c_{id} should be a measure of *perceived* distance, or possibly journey time. However, no such measures were available for this study, and in any case such refinement is probably unnecessary. To avoid absurdly high weights being placed on very small distances, it is usual to add a constant distance - known as an "intrazonal cost" - to each calculated distance. Finally, possibly the most troublesome aspect of modelling is the choice of deterrence function f(.). Scrutiny of the spatial location literature suggests a wide range of possible functional forms.

4.29 The study team experimented with a range of deterrence functions and intra-zonal costs, and examined the results in detail with their technical advisers. The accessibility scores were found to be reasonably robust to choice of deterrence function, intrazonal cost and measure of attractiveness. It was eventually decided to use an inverse square deterrence function and an intrazonal cost of 10km, and to measure attractiveness by the average number of available beds in the acute sector. When long stay specialties were modelled separately, the number of beds in long stay specialties was substituted. In the acute sector, this resulted in an accessibility index with a national average of 2.34 and a range of 0.53 to 4.75. The lowest values occurred in rural parts of Cornwall and Northumberland. The highest values were found in London and the Tyne and Wear conurbation.

4.2.2 Measuring accessibility of GP services

4.30 The extent to which hospital services are used by individuals may also be determined by the level of provision of primary care services. The range of primary care services provided by the NHS is extensive and the associated data are poor, making measurement of supply difficult. However, the supply of primary care can be reflected by the provision of General Practitioner services. Indeed, in their role as 'gatekeepers' to the health care system, GPs may have a large impact on utilization. Two competing hypotheses can be formulated:

(1) A relatively high provision of GPs in a geographical area will result in a relatively high hospital utilization rate, as more individuals will gain entry to the health care system (assuming there is unmet demand for health care). Consequently, under this hypothesis, more individuals are referred to the hospital sector as GP provision increases.

(2) A relatively high provision of GPs in an area will result in a relatively low rate of hospital utilization, as GPs provide services which are substitutes for some services provided in the hospital sector.

In practice, of course, both phenomena may operate. However we can only observe the net effect of the two. Hypothesis 1 is probably dominant in an area with a supply of hospital services that is high in relation to 'need'; and hypothesis 2 is probably dominant in areas in which hospital supply is relatively low. In the latter areas of suppressed utilization, poor people might have unmet needs for hospital care, richer people might use more private care, and both might make high use of primary care. Whilst, in principle, phenomena such as these imply that we should investigate the pairwise interactions between supply variables, time constraints made this impossible.

4.31 The impact of GP supply on hospital utilization is determined first through individuals' perceptions of the availability of GP services, and then through GP perceptions of hospital provision. Hospital provision is dealt with in the preceding Section. Perceived availability of GP services to individuals will be affected by a multitude of factors, many of them determined by individuals' personal characteristics and attitudes, and by their socio-economic circumstances. Attempting to model characteristics and attitudes meaningfully is beyond the scope of a macro study such as this, and socio-economic circumstances were dealt with elsewhere in this study (Section 4.1). We nevertheless sought to capture some aspects of GP provision which are likely to be salient to individuals. As with hospital provision, this requires the reconciliation of quantity, distance and competition from neighbouring populations into a single index.

4.32 To this end, the study team had available the address of the principal surgery of all practices in England, from which could be inferred the surgery's grid reference. The number of doctors employed was the only measure of quantity. Clearly this is a very limited database. Ideally we should like to also like to incorporate measures of the *quality* of primary care provision. However, no comprehensive measures of list size, practice nurse provision, branch surgeries or other quality surrogates were available. In the absence of a more comprehensive data set, accessibility to GP services was therefore measured in a similar fashion to accessibility to hospitals. The measure of provision used was the number of GPs employed, and distance and deterrence were modelled as in Section 4.2.1.

4.33 The interpretation of the GP access variable caused the study team and its advisors some difficulty. For example, it might plausibly be argued that higher levels of GP provision are positively associated with indicators of poor quality of primary care, such as a high proportion of single handed GPs over 65. It is therefore possible that, at least in part, the variable might reflect shortcomings in primary care in an area rather than the level of provision. We nevertheless felt it was

important to include the variable in our models in order to seek to capture the relationship between primary care provision and hospital utilization. This is clearly an area for further more detailed research.

4.34 The GP accessibility index had a mean of 0.53, and a range of 0.16 to 0.96. There was considerably more geographical dispersion of the low and high values amongst wards than was found with the NHS hospital index. Rural areas featured amongst both the highest and lowest scoring wards, possibly reflecting the fact that we could not model branch surgeries. In general, wards close to the main surgery are likely to be given a relatively high score and those close to a branch surgery correspondingly low scores. London wards tend to exhibit high levels of GP accessibility, reflecting the relatively short distances to GP surgeries found in the conurbation.

4.2.3 Measuring provision of nursing and residential homes

4.35 The extent of provision of nursing and residential home facilities in a particular ward may have significant effects on the hospital utilization rate of that ward. Two hypotheses require investigation:

(1) Nursing home beds are substitutes for hospital beds. Therefore wards with relatively more provision of nursing home facilities will exhibit relatively lower utilization of hospital services, if all other factors are held constant. This may be because the care provided by the home obviates the need for an admission, or because home residents are discharged from an episode earlier than other patients.

(2) Nursing home beds complement hospital beds. Therefore wards with relatively higher provision of nursing home facilities will have relatively higher utilization of hospital services, if all other factors are held constant.

This may arise because those in nursing homes are more likely to have their problems diagnosed. This implies that nursing homes only provide care which is not available in the hospital system.

4.36 In practice it may be true that nursing homes beds behave both as substitutes and complements to hospital beds, and the magnitude of each individual effect would be difficult to establish. Nevertheless, it may be possible to estimate the net effect of nursing home provision on hospital utilisation rates.

4.37 Data regarding the provision of public and private sector nursing and residential home facilities at ward level are available from the 1991 Census of Population, which gives details of the number of residents by broad age category. Clearly these data give no information about the quality of provision. However, they are comprehensive and reliable, so we chose to use the proportion of the population of the ward aged 75+ resident in such homes as the relevant index of supply. It is possible to construct more complex indices. However, such refinements seem unnecessary, as this ratio is likely to capture the order of magnitude of supply of nursing and residential homes. Thus, in contrast to the other supply variables, the supply of homes is measured in terms of the absolute level of provision within the ward, and takes no account of provision in neighbouring wards.

4.2.4 Measuring accessibility of private hospitals

4.38 The use made of private health care by individuals clearly may have an impact on utilization of NHS facilities. Much private care is purchased through insurance arrangements, the prevalence of which is likely to be dependent principally on the socio-economic considerations modelled elsewhere. Thus the study methodology implicitly takes account of the predisposition of different areas to purchase different levels of private health care. However, it is also possible that - in addition to socio-economic determinants - location has an impact on the use made

of private health care facilities. The provision of hospital services in the independent sector therefore also merits attention.

4.39 Again, two hypotheses can be explored: that private hospital beds are substitutes for NHS hospital beds; and that private hospital beds are complements for NHS hospital beds. Both effects might appear in practice and it would be difficult to identify them individually. The 1991 Census of Population gives the number of visitors present in private hospitals on Census night by ward, and implied a total of 8,524 occupied non-psychiatric beds throughout England. Clearly not all such visitors are necessarily patients. And the number of patients on Census night might be a very imprecise measure of total capacity of the hospital. However, we found that the comparable Census measure for NHS hospitals does give an accurate reflection of total bed availability. As a result, we feel that the Census offers an acceptable measure of private inpatient provision.

4.40 Accessibility to private facilities was therefore modelled as in Section 4.2.1 (accessibility to NHS hospitals), using the same deterrence function and intrazonal costs. It did not seem necessary to separate private supply between acute and long stay. The number of visitors in non-psychiatric private beds was the measure of attractiveness used, and the measure of location was the grid reference of the zone centroid. Although this is not as precise as postcode grid reference, it is considered adequate for our purposes.

4.41 The private hospital accessibility index has a mean of 0.17 and a range of 0.02 to 2.07, indicating a distribution skewed to the left, with a small number of very high scoring wards.

4.2.5 Higher level supply variables

4.42 The focus of this study is the use made by small areas of NHS inpatient facilities.

The supply variables described above therefore seek to represent supply considerations at the small area level. However, for two reasons, many characteristics of supply cannot be satisfactorily modelled at such a parochial level. First, although the level of a phenomenon may vary between small areas, it may be infeasible to develop a measure of the phenomenon at ward level. The relevant data may therefore be available only at a higher level of aggregation. And second, some phenomena are truly invariant across the whole of a DHA or FHSA (or indeed RHA). In particular, policy variables may operate on all wards within an administrative area.

4.43 Through the use of multilevel modelling, as described in Section 5.2, the study methodology permits incorporation of variables specified at a level higher than the synthetic ward (Paterson and Goldstein, 1991). Many such data are available from the Health Service Indicators provided by the NHS Management Executive (1992). Most of the high level variables tested were therefore derived from this source. Where these were considered to be dependent on local socio-economic conditions, the variables were first regressed on a series of socio-economic variables to derive "expected" levels. The multilevel analysis then used the ratio of actual to expected as a measure of relative local provision. The variables tested were as follows:

District health authorities

Staff per capita: medical and dental
Staff per capita: nursing and midwifery
Staff per capita: administrative & clerical
Staff per capita: ancillary
Teaching district (yes/no)
Percentage of expenditure on estate management
Percentage of buildings in condition C or D
Percentage of theatre sessions cancelled
Ratio of decisions to admit to GP referrals

Family health service authorities

Expenditure per head
Percentage of practices below minimum standards
Percentage of GPs with list size < 1,000
Percentage of GPs with list size > 2,500.
Percentage of GPs under 65 in single-handed practices
Percentage of practices without a nurse

4.3 Measuring the utilization of NHS resources

4.44 The purpose of this study is to devise a formula for equitably distributing finances for HCHS to geographical areas. The formula should indicate the costs to the NHS of providing the *national average level of care* to a locality, given the social and health characteristics of the area, and assuming *national average policies and levels of efficiency*. To that end, the empirical work should seek to link social and health conditions to utilization measured in terms of costs to the NHS. However, identifying the true costs to the NHS of particular patterns of utilization is difficult. RAWP used inpatient bed days and the Review used episodes as proxies for the total revenue consequences of utilization. In this Section we first describe the dataset from which utilization measures were derived, and then explain how the measures were calculated.

4.3.1 The Hospital Episode Statistics

4.45 The measures of utilization available for this study were derived from an abstract of the 1990/91 Hospital Episode Statistics (HES), a database of hospital inpatient discharges and deaths made available to the study team. The HES abstract contained the following information:

Method of admission
Source of admission
Category of patient (NHS vs private)
Wait for elective admission
Age group
Specialty group

Operation group (x4)
Order number of episode
Episode duration
Discharge destination
Patient classification (day vs ordinary case)
Synthetic ward of residence

4.46 The HES extract contained an indicator of the synthetic ward of residence. Utilization rates for each ward could therefore be found by relating total utilization in a ward to the mid-1991 population estimates for the ward.

4.47 The HES system was introduced in April 1987. It covers all specialities and is based on finished consultant inpatient and day case episodes (an episode where the patient has completed a period of care under a consultant and is either transferred to another consultant, is discharged, or dies). In addition, in some Regions, the HES contain information on unfinished episodes. However, coverage was not comprehensive enough to be useful for this study, so only completed episodes were analyzed. The HES system is based on treatment in hospitals and therefore does not cover outpatient or community activity. The data for HES are collated in the Hospital Patient Administration system and are submitted via the District and Regional information systems to the OPCS. The HES data set used in this study contained some 8,566,887 valid records, covering all finished inpatient and day case hospital episodes for the financial year 1990/91. Data for 1991/92 were also made available towards the end of the study.

4.48 It should be noted that there is some variation amongst DHAs in the interpretation of what constitutes a finished consultant episode. There is a degree of flexibility in its definition which means that comparisons in episode rates between Districts may not be strictly valid (Clarke and McKee, 1992). However the extent to which this might result in systematic inter-District biases is unknown. The multi-level modelling strategy we use seeks to correct for systematic inter-District variation.

4.49 Inevitably, such a large data set - constructed from a variety of independent sources - is unlikely to be error free. Completeness of the data was assessed by comparison with the statistical return KP70. This provides annual totals for finished consultant episodes ending during the financial year and is completed independently of HES. For the year being studied, we were advised that the count of episodes from KP70 was likely to be more reliable than that from HES. For most Regions, the KP70 return was within 5% of the number of episodes recorded in HES. The exceptions were North East Thames (HES 28% higher than KP70) and North West Thames (HES 6% lower than KP70). In constructing the utilization measures employed in this study, each episode was multiplied by a factor defined as the ratio of the number of completed episodes from the KP70 return to that derived from the HES system for the District in which the episode took place (Department of Health, 1993). The HES contained no data relating to patients treated in Rugby HA. We were therefore forced to omit all wards within the catchment of Rugby HA from the analysis.

4.50 We sought more detailed advice on the likely accuracy of the HES data and were made aware of a number of common coding errors (such as entering the date of birth where the date of admission is required). It was also suggested that coding of operations and waiting times might be especially weak. We were however unable to make any adjustments for these shortcomings. The one systematic and potentially significant error which we examined in some detail concerned the imputation of an incorrect dummy postcode where the correct postcode was not known. There is evidence that, in certain Districts, clerks have allocated all inpatients with unknown postcodes to a "dump" postcode within the District. This results in a very large number of episodes being assigned to the ward in which the postcode is found. A scrutiny of heavily used postcodes gave some indication of large scale dumping, and we have omitted the associated wards from the analysis. Note that, where dumping occurred, there was likely to be a concomitant under-recording of utilization in neighbouring wards. We could not adjust for this

phenomenon, except to delete a very small number of wards with exceptionally low utilisation rates. A total of 45 of the 4985 wards were omitted from the analysis.

4.51 There were no costs attached to the HES database. Any costs we were to use therefore had to be inferred from a very limited amount of information available on each inpatient episode, namely: length of stay; specialty group; age group; sex. The twelve specialty groups were determined by the Management Executive in advance of commissioning the study. They comprise very broad groupings of medical specialty, as shown in Table 4.3.

4.52 Table 4.4 describes the basic characteristics of the twelve specialty groups. The largest numbers of episodes occur in surgery and medicine. Lengths of stay varied enormously between specialty groups, and the largest numbers of bed days were in psychiatry and mental handicap. However, these data should be viewed with some caution. In the year being studied (1990/91) a large number of long stay inpatients were discharged into the community. In a steady state, the use made by discharged patients should be a proxy for total use of inpatient facilities. However, when the number being discharged is larger than the number being admitted, the total use of bed days may be overstated. As a result, the numbers shown here may overstate the "steady state" use of inpatient facilities in long stay specialties - notably geriatric medicine, psychiatry and mental handicap. In order to emphasize the scale of this problem, Table 4.4 gives average lengths of stay and the proportion of bed days attributable to patients whose stay exceeded one year, demonstrating the huge differences between specialties.

Specialty group	Description	Specialty codes
1. Surgery	All surgery excluding neurosurgery, plastic surgery, cardiothoracic surgery and paediatric surgery	100-199 (not 150,160,170,171)
2. Surgery II	Neurosurgery, plastic surgery, cardiothoracic surgery and paediatric surgery	150,160,170,171
3. Medicine	All medical excluding geriatrics, cardiology, medical oncology, neurology	300-450 (not 430,320,370,400)
4. Geriatrics	Geriatric medicine	430
5. Medicine II	Cardiology, medical oncology, neurology	320,370,400
6. Psychiatric	Psychiatric	710-715
7. Mental handicap	Mental handicap	700
8. Maternity	Maternity	501,610
9. Gynaecology	Gynaecology	502
10. Radiotherapy	Radiotherapy & radiology	800,810
11. Other	All other valid episodes with code	620,820,901
12. Not stated	All episodes invalid code	-

Table 4.3: Definitions of specialty groups

Specialty group	Percentage of all episodes	Percentage of all bed days	Average length of stay (days)	Percentage of episodes with zero stay (%)	Bed days from stays > 1 year (%)	Elective episodes as % of all	GP fund-holding episodes (%)
1. Surgery	35.1	14.6	4.5	28.8	7.7	66.9	22.1
2. Surgery II	2.7	1.3	5.5	26.6	3.2	66.6	12.7
3. Medicine	29.0	15.0	5.6	19.6	8.4	20.5	6.5
4. Geriatrics	5.2	14.9	31.7	3.3	32.6	14.3	1.0
5. Medicine II	2.1	1.4	7.5	20.5	18.6	60.1	2.5
6. Psychiatric	2.4	25.1	117.5	3.1	71.4	20.6	0.2
7. Mental handicap	0.5	19.5	401.1	2.8	97.4	80.0	0.0
8. Maternity	9.9	3.1	3.5	13.6	3.1	4.2	1.1
9. Gynaecology	10.0	2.4	2.6	32.2	4.3	61.8	31.1
10. Radiotherapy	0.9	0.6	7.2	23.2	11.4	76.8	1.2
11. Other	1.7	1.8	11.6	15.8	13.2	35.9	1.5
12. Not stated	0.4	0.3	8.3	15.8	16.9	27.1	11.1
ALL SPECIALTIES	100.0	100.0	11.1	22.3	45.0	42.1	13.4

Table 4.4: Descriptive statistics derived from the HES abstract, 1990/91

4.53 In addition, Table 4.4 shows the extent of elective (as opposed to emergency) admissions. The average of 42% again disguises big differences between specialty groups. A subset of elective admissions is those admitted for at least one of the procedures included in the list purchased by GP fundholders. Such episodes accounted for 13% of all admissions, predominantly in the three surgical specialty groups.

4.3.2 Calculating utilization measures

4.54 All utilization variables were standardized (using indirect standardization) for age and sex, yielding a ratio of actual to expected utilization in a ward. Four types of measure of utilization were calculated:

(a) number of episodes;
(b) number of bed days;
(c) estimated costs;
(d) standard costs.

4.55 Clearly, the *number of episodes* is insensitive to case mix and severity, and may vary depending on local policy regarding transfers between consultants and readmissions. To some extent, the *number of bed days* accommodates the problem of differences in severity by weighting each episode according to number of bed days occupied. However, this index attaches a weight of zero to day cases, and makes no allowance for differential costs between specialties. In addition, it can be distorted by a small number of very long lengths of stay, and, as noted above, 45% of bed days from the 1990/91 HES related to episodes lasting more than one year. When modelling this measure of utilization we ignored all bed days in excess of one year.

4.56 The *estimated costs* of episodes are calculated by attaching specialty-specific fixed

costs to each episode, and adding specialty-specific variable costs for each day's stay. This measure therefore combines the first two measures of utilization, and allows a more sensible treatment of day cases. Furthermore, it captures variations in intensity of use brought about by differences in patient characteristics. However, variations in lengths of stay (and therefore estimated costs) may also reflect differences in levels of efficiency and case management between areas, unrelated to health care needs. In addition, there are considerable problems involved in arriving at satisfactory measures of fixed and variable costs.

4.57 The most serious problem relating to estimated costs is however their vulnerability to distortion by very long lengths of stay. As noted above, we had available only data pertaining to finished episodes. Ideally, we would have wished to measure resource use relating to finished and unfinished episodes for the year studied, regardless of when the patient was admitted or discharged. However, because we had available length of stay data, and neither admission nor discharge date, we were unable to determine which portion of resource use fell within the period. The analysis of finished consultant episodes on which we rely can only be a proxy for resource use in the period studied. Also, the measures of ward utilization can be seriously distorted by a single very long length of stay, so when testing the estimated costs variable we again truncated episodes lasting more than one year to just one year. Clearly the choice of this cut-off point is arbitrary, but one must be chosen to prevent the utilization variable being dominated by a relatively small number of long stay episodes.

4.58 *Standard costs* are defined as the national average costs for the age, sex and specialty group associated with an episode. They effectively offer a count of episodes, with each episode weighted according to the national average level of use according to the specialty group and the patient's age and sex. They are therefore still dependent on identification of satisfactory cost structures for each specialty. In addition, they may obscure the effects on lengths of stay of

variations in severity or social needs between areas not captured by age and sex. However, by assuming a national average use of resources, they overcome the problem of differences in local levels of efficiency, and they are less sensitive than actual costs to long lengths of stay. Moreover, by measuring resource use assuming an average level of efficiency and medical practice, use of standard costs appears to be in line with the spirit of the resource allocation philosophy adopted in this and previous studies.

4.59 The study team and its advisors spent a considerable amount of time debating the relative merits of the four measures of utilization, and preliminary modelling work was undertaken using all four types of measure. Clearly - providing that we could find reasonably acceptable measures of fixed and variable costs - estimated costs and standard costs were likely to be superior to bed days or episodes as measures of the resource consequences of utilization. Because of the complexity of the modelling procedure and the short time scale of the project, we were obliged to focus on one measure to develop a provisional model. In consultation with our advisers, we therefore chose to focus on estimated costs as the measure of utilization in the acute sector, in the knowledge that this might be distorted by variations in efficiency, if these are systematically related to socio-economic circumstances. However, the ability of estimated costs to capture variations in intensity of use brought about by variations in need (other than that due to age and sex) was considered important enough in the acute sector to outweigh any shortcomings. We nevertheless also examined the implications for the chosen model of substituting standard costs as the measure of utilization.

4.60 In the non-acute sector there is a very high preponderance of episodes with long lengths of stay. The use of estimated costs is therefore less satisfactory. In mental handicap, mental illness and psychiatry we therefore focused on standard costs, again in consultation with our advisers.

4.61 In the calculation of costs, a fixed episode cost and a variable bed days cost must first be specified for each specialty group. Although several potential sources of cost data were explored, the only cost data that the Department of Health were able to make available to the study team was a statistical analysis of the relationship between episodes, bed days and hospital costs at the national level undertaken by East Cheshire Statistical Analysis Consultancy. These calculations are based on the assumption that fixed and variable specialty costs are constant across age groups. Whilst this is unlikely, there was no alternative data source which satisfied this study's requirements. The results are summarized in Table 4.5.

Specialty group	Cost per episode (£)	Cost per bed day (£)
1. Surgery	153.3	165.1
2. Surgery II	253.8	277.1
3. Medicine	149.5	123.6
4. Geriatrics	0.0	98.8
5. Medicine II	560.3	111.8
6. Psychiatric	1031.3	86.5
7. Mental handicap	1031.3	85.9
8. Maternity	144.1	147.3
9. Gynaecology	144.1	153.7
10. Radiotherapy	144.1	159.0
11. Other	144.1	91.8
12. Not stated	181.2	119.8

Table 4.5: Episode and bed day costs by specialty group
(Source: East Cheshire Statistical Analysis Consultancy)

4.62 By applying the costs in Table 4.5 to the number of episodes and number of bed days in each specialty group associated with a ward, the estimated costs of

inpatient utilization in the ward can be derived. However, in order to calculate the estimated costs dependent variable, a measure of the *expected* costs in the ward must be calculated.

4.63 The next stage in the calculation of costs was therefore to identify, for each specialty, *national* rates of admission and lengths of stay by age and sex. The product of these two variables gives the number of bed days per head of population. The cost data for the relevant specialty are applied to episode and bed day rates to yield a national average cost per head. By way of illustration, Table 4.6 gives utilization by age and sex for all acute specialties (therefore excluding maternity and the three long stay specialties: geriatric medicine, psychiatric and mental handicap). The Table gives summaries of per capita admission rates; average lengths of stay per episode; per capita bed days; per capita costs; and costs per episode. It should be noted that these data include the healthy newborn, who are in most Districts reported as episodes in specialty group 3. Table 4.7 reports the same data for the three long stay specialties in aggregate.

4.64 The expected costs in a ward for each specialty, age and sex group are therefore defined as the product of the national average costs per head and the population in the relevant age/sex group. These were summed to give total expected costs in the ward. The age/sex standardized measure of the estimated costs dependent variable was then defined as estimated costs divided by expected costs.

4.65 The standard costs of an episode are defined as the national average costs for an episode for the relevant age, sex and specialty group. Standard costs could therefore be inferred from the age/sex/specialty specific average cost per episode in the national analysis, permitting calculation of the standard costs incurred by each ward. These were divided by the ward's expected costs, as defined above, to yield a standardized measure of standard costs. Utilization measures were calculated for each specialty group separately, as well as for all specialties.

4.66 The standardized utilization measure was scaled to have a national mean of 100. For all acute specialties (excluding maternity) expected costs were found to have a standard deviation of about 25. Chapter 5 explains how we sought to explain variations in utilization measures. For most of the analysis, the newborn were retained in the analysis. It was found that excluding the newborn made very little difference to results.

Age	0-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-85	85+	All
EPISODES PER 1000 POPULATION																			
Male	379.0	95.9	69.4	72.5	76.9	74.3	78.7	83.3	83.6	95.9	127.5	161.8	203.4	247.7	274.4	319.0	317.0	323.0	140.8
Female	322.6	67.9	59.3	113.3	146.0	150.0	140.2	129.0	122.0	126.6	140.5	142.2	154.5	177.2	184.1	210.2	211.5	210.4	150.8
All	351.5	82.3	64.5	92.3	110.7	111.6	109.3	106.1	102.8	111.2	134.0	152.0	178.2	209.9	223.2	252.8	246.8	237.4	145.9
LENGTH OF STAY																			
Male	3.72	2.42	3.04	3.58	3.27	3.56	3.13	3.60	3.85	4.56	5.08	5.47	6.02	6.50	6.96	7.84	8.73	10.06	4.93
Female	3.79	3.13	3.00	2.56	2.33	2.45	2.86	3.13	3.70	4.04	4.38	5.48	6.40	7.22	8.23	9.58	11.25	13.82	5.03
All	3.75	2.71	3.02	2.97	2.66	2.83	2.96	3.32	3.76	4.27	4.71	5.48	6.19	6.82	7.56	8.72	10.17	12.60	4.98
BED DAYS PER 1000 POPULATION																			
Male	1410.5	232.4	211.4	259.7	251.4	264.7	246.2	300.0	321.7	437.8	647.7	885.6	1223.9	1609.7	1910.8	2500.0	2766.7	3250.7	694.3
Female	1222.9	212.6	177.7	290.3	339.9	368.2	401.5	404.1	450.8	511.6	615.3	778.9	989.2	1278.3	1515.8	2014.1	2379.9	2908.2	758.4
All	1319.2	222.8	195.0	274.6	294.7	315.7	323.4	351.9	386.2	474.6	631.5	832.0	1102.6	1432.1	1686.7	2204.3	2509.5	2990.3	727.1
COSTS PER HEAD (£)																			
Male	249.24	53.20	44.73	53.78	53.01	53.55	51.34	59.62	63.55	82.96	121.15	160.98	216.65	279.41	323.93	411.52	443.71	503.27	125.5
Female	211.31	43.30	37.35	62.44	74.20	78.92	82.32	81.58	88.04	98.41	115.75	139.07	171.05	216.20	248.78	322.82	370.97	436.96	134.1
All	230.78	48.39	41.15	57.99	63.39	66.04	66.73	70.58	75.79	90.67	118.45	149.97	193.09	245.53	281.29	357.55	395.33	452.84	129.9
COSTS PER EPISODE (£)																			
Male	657.7	554.9	644.3	741.9	689.0	721.0	652.1	715.8	760.0	864.7	949.9	995.0	1065.2	1127.9	1180.6	1290.1	1399.5	1558.0	891.3
Female	655.0	638.2	629.6	551.2	508.3	526.0	587.2	632.6	721.9	777.5	823.6	977.8	1106.8	1220.3	1351.2	1535.5	1754.2	2076.6	889.5
All	656.5	588.2	637.7	628.2	572.3	591.9	610.7	665.3	737.5	815.2	883.7	986.9	1083.8	1169.7	1260.5	1414.3	1601.6	1907.6	890.4

Figure 4.6: Utilization rates by age and sex, all acute specialties, 1990/91

Age	0-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-85	85+	All
EPISODES PER 1000 POPULATION																			
Male	0.6	2.0	2.8	4.7	6.7	6.9	6.4	6.2	5.4	4.9	4.6	4.6	6.2	15.1	34.0	84.5	145.5	220.2	12.3
Female	0.6	1.2	2.1	4.1	6.1	5.9	5.8	6.0	5.4	5.4	5.5	5.5	6.8	14.6	30.8	72.5	124.5	182.3	16.6
All	0.6	1.6	2.4	4.4	6.4	6.4	6.1	6.1	5.4	5.2	5.1	5.0	6.5	14.9	32.2	77.2	131.6	191.4	14.5
LENGTH OF STAY																			
Male	973	547	403	380	150	152	148	117	104	111	110	102	84	50	37	31	30	31	83
Female	782	774	468	356	194	168	168	182	140	128	133	124	112	68	53	43	43	42	79
All	886	628	429	369	170	159	157	149	122	120	122	114	99	60	45	38	38	39	81
BED DAYS PER 1000 POPULATION																			
Male	623	1121	1130	1768	1001	1050	950	721	561	550	509	464	521	763	1244	2609	4296	6719	1019
Female	439	936	970	1457	1174	991	979	1091	758	690	728	684	759	998	1625	3144	5301	7696	1319
All	534	1031	1052	1617	1086	1021	965	905	659	620	618	574	644	889	1461	2934	4964	7462	1173
COSTS PER HEAD (£)																			
Male	54	98	100	157	93	98	88	68	54	52	48	44	50	74	123	258	426	668	95
Female	38	82	85	130	107	91	90	100	71	65	68	64	72	96	158	307	519	759	124
All	47	90	93	144	100	95	89	84	62	59	58	54	61	86	142	288	488	737	110
COSTS PER EPISODE (£)																			
Male	84714	48045	35652	33668	13932	14095	13766	11090	9915	10578	10372	9678	8089	4919	3603	3052	2930	3034	7773
Female	68298	67484	41231	31603	17694	15499	15495	16714	13101	12073	12435	11721	10517	6580	5114	4240	4169	4161	7465
All	77264	55000	37944	32730	15683	14731	14583	13859	11505	11357	11489	10800	9401	5796	4424	3731	3710	3850	7592

Table 4.7: Utilization rates by age and sex, lóng stay specialties, 1990/91

4.4 Conclusion

4.67 The data used in this study were derived from a variety of sources. The measures described in this Chapter fall into three categories: needs indicators, supply variables and utilization variables. Their immediate purpose was to model the determinants of utilization in small areas, as described in Chapter 5. However, the database assembled for this study is also clearly relevant to exploring a wide range of research and policy issues, and we are glad that it will be made available to health service and research workers. It represents a considerable improvement on any data set previously used in the UK for this kind of analysis. In particular, in comparison with the previous Review of RAWP:

(a) We have been able to incorporate data for almost the whole of England.

(b) In measuring possible indicators of needs, we were able to include data on limited long standing illness and low birth weight at the ward level. Equally, we benefited from Census variables which were contemporaneous with the HES and vital data.

(c) We have been able to take advantage of the far greater use made of postcoding to identify the location of provider units in both primary and hospital care. This has allowed us to develop a more comprehensive consideration of supply. In addition, in measuring supply, we have been able to incorporate Health Service Indicators into the analysis.

(d) In measuring utilization, we did not need to rely on a crude count of episodes. Instead we were able to develop a direct measure of resource use, broken down by specialty grouping.

(e) We have been able to incorporate psychiatric episodes into the analysis.

References

Clarke A and McKee M (1992) The consultant episode: an unhelpful measure, *British Medical Journal,* 305, 1307-1308.

Department of Health (1993) *Hospital Episode Statistics: Volume 1 England, 1989/90,* The Department, London.

Lubitz J and Prihoda R (1982) The use and costs of health care in the last two years. Self report morbidity shows social variations in response of life, *Health Care Financing Review* 5, 117-131.

Mays N and Bevan G (1987) *Resource Allocation in the Health Service*, Occasional Paper in Social Administration 81, Bedford Square Press, London.

National Health Service Management Executive (1992), *Health Service Indicators Handbook*, Department of Health, London.

O'Donnell O and Propper C (1991) Equity and the distribution of UK National Health Service resources, *Journal of Health Economics* 10,1-19.

Paterson L and Goldstein H (1991) New statistical methods for analysing social structures: an introduction to multilevel models, *British Educational Research Journal*, 17(4), 387-393.

Sheldon T A, Davey Smith G and Bevan G (1993) Weighting in the dark: resource allocation in the new NHS, *British Medical Journal* 306, 835-839.

5. MODELLING NHS INPATIENT UTILIZATION

5.1 Chapter 3 set out the theoretical model of demand for health care on which this study is based. Chapter 4 described the data at small area and higher area levels that were available to build an empirical model of demand. This Chapter describes how the empirical model was made operational. Section 5.1 gives the basic modelling strategy used to estimate the model at small area level. Section 5.2 describes how the model results can be used to derive a resource allocation formula, and Section 5.3 gives an outline of the multilevel modelling methods used to incorporate supply variables specified at DHA and FHSA level. More technical discussion of the modelling process and the use of multilevel methods are given in Appendices D and E.

5.1 Estimating a small area model of utilization

5.2 Chapter 3 explained that a fundamental requirement is to estimate an equation of the form:

$$U_i = g_1(N_i, S_i) \qquad (5.1)$$

where U reflects utilization, N indicators of health care needs and S perceived supply. The statistical problem that arises is that - for each of the perceived supply variables S - there may be operating a simultaneous relationship of the sort:

$$S_i = g_2(N_i, U_i, X_i) \qquad (5.2)$$

where X represents broader socio-economic considerations not directly related to the need for health care. In other words, supply itself is a function of health care needs, other socio-economic factors and utilization.

5.3 Econometricians call this phenomenon endogeneity, and the variables U and S are termed endogenous variables, in the sense that they are determined within the system of equations. In contrast, the needs and socio-economic variables N and X

are exogenous, in the sense that they are determined outside the equations. Endogeneity is commonly found when modelling economic phenomena. It gives rise to problems when seeking to develop empirical estimations of equations such as (5.1) because, if ordinary least squares (OLS) regression is used, the variable S will not in general be independent of the residuals. This breaches one of the fundamental assumptions of statistical modelling, and may lead to biases in the estimated coefficients of the model, and therefore a faulty formula. Econometricians have developed analytic techniques to test for endogeneity, and to correct for it when it is found. This study has made use of these techniques.

5.4 A major problem confronting the study team was that it had no incontrovertible *a priori* reason for excluding *any* of the socio-economic variables described in Section 4.1 from possible consideration as one of the "needs" variables N. The methodology we describe therefore gave *every* socio-economic variable the chance of being considered as a needs variable. However, to include indiscriminately every socio-economic variable in the equation would not be helpful, as many of the variables were highly correlated with supply, so it would be impossible to disentangle needs effects from supply effects, as required. Therefore, as we explain later in this Section, only those socio-economic variables that proved statistically important were eventually retained in the set N. The remainder were placed in the set X of "other" socio-economic variables.

5.5 In order to estimate equation (5.1) above, the following strategy was adopted, assuming a set of needs variables N had been identified. The statistical representation of (5.1) was specified in the usual linear form:

$$U_i = \alpha + \sum_{j=1}^{m} \beta_j N_{ij} + \sum_{j=1}^{n} \gamma_j S_{ij} + \epsilon_i \qquad \textbf{(5.3)}$$

where U_i is some measure of utilization in synthetic ward i; there are m needs indicators N and n supply variables S; α, β and γ are vectors of parameters to be

estimated; and ϵ is the usual error term.

5.6 The functional form of the equation must be chosen. We considered two alternatives: an additive model and a multiplicative model. In the additive model, variables are used as described in Chapter 4. In the multiplicative model, utilization is modelled as follows:

$$U=a\prod_{j=1}^{m} N_j^{b_j}\prod_{j=1}^{n} S_j^{c_j} \qquad \textbf{(5.4)}$$

where a, **b** and **c** are vectors of parameters. In order to make this model operational, the natural logarithm of all variables must be taken before estimating equation (5.3). It is noteworthy that, if the multiplicative model is used, then the parameters c_j satisfy the relationship

$$c_j=\frac{S_j}{U}.\frac{\partial U}{\partial S} \qquad \textbf{(5.5)}$$

That is, the coefficient of $\log_e S_j$ estimated in the utilization equation indicates the *elasticity* of utilization with respect to the supply variable j. It can be interpreted as the percentage increase in utilization brought about by a 1% increase in supply S_j. Similarly, the coefficient on each of the health needs variables indicates the elasticity of utilization with respect to the associated needs variable.

5.7 Although we tested both additive and multiplicative models, we found that, in terms of specification, the multiplicative model performed consistently better than the additive, and therefore, unless stated otherwise, it should be assumed that the multiplicative model is being used. This is the same functional form that was used in the Review of RAWP.

5.8 Having chosen the form of the model, a standard statistical test was undertaken to test for endogeneity of the supply variables S. This entailed regressing each of S

variables on a set of "instruments", which in this context can be considered to be possible socio-economic determinants of supply. In addition to the needs variables N which were to be included in the final equation (5.4), a wide range of other socio-economic variables X was included in the list of instruments, as described in Appendix B. The residuals from these preliminary regressions (the variation in S not explained by the instruments) were entered as additional explanatory variables into equation (5.4), which was then estimated using OLS. If using a standard F-test these residuals were found to be jointly statistically significant in modelling utilization, then the supply variables were considered to be endogenous, and OLS could therefore not be employed to estimate equation (5.4).

5.9 Instead, the method of two-stage least squares was used (see Appendix D). This is a standard econometric technique, employed when endogeneity is present. It entails first regressing each of the supply variables S on the set of instruments as above, and then using the *predicted* values of the supply variables as explanatory variables in the regression, in place of their actual values. This approach should then yield consistent estimates of all coefficients in the regression of utilization on needs and supply. Standard errors are adjusted to take account of the endogeneity. In practice, we found that the supply variables were indeed endogenous, and that a method such as two stage least squares (2SLS) was therefore necessary. It should be emphasized that although the instrumented supply variables are used in estimating all the coefficients in the utilization equation, the coefficients so estimated apply to the original supply variables and not to the instrumented supply variables.

5.10 After estimating the utilization equation, it was necessary to test that the model was well specified: that is, that the model conformed to all the requirements of the econometric technique employed. For example, it was necessary to test whether there were problems with the functional form, or important omitted variables. The specification test used is described in Appendix D. It should be noted,

however, that with the large number of observations used in this study it was almost inevitable that we would have the power to detect some element of misspecification. We nevertheless sought to keep the extent of such misspecification modest.

5.11 Furthermore, we were interested in finding as parsimonious a model as possible: that is, a model with the least number of variables which sensibly capture variations in supply-adjusted utilization. As well as being good statistical practice, this results in a formula which is not too large or complex. In order to obtain a parsimonious model, we employed a "general to specific" methodology, in which in the first instance a large number of possible needs indicators N were used. This gave rise to what is known as the unrestricted model. By setting the coefficients on certain N variables equal to zero, we effectively removed them from the model, and therefore moved to a restricted model. Tests were undertaken to ensure that such restrictions were valid. In addition, it was necessary to test other possible sources of misspecification, such as endogenous instruments and heteroscedasticity. The various tests used in this study are described in Appendix D.

5.12 Throughout, each observation was weighted in proportion to the total population of the ward. This ensured that, in seeking to infer a national average model of utilization, we did not give undue weight to patterns of utilization in smaller synthetic wards.

5.13 In practice, the large number of variables available to the study team required a modelling methodology to enable us to discard all but the most important variables before starting the "general to specific" process. We therefore developed the following strategy. First, as explained above, all health status and socio-economic variables were specified as instruments. Then a "minimal" model was estimated, comprising the four endogenous supply variables. The residuals from this model

were correlated with the instruments, and the instrument exhibiting the highest correlation was added to the set N of explanatory needs variables. In this way, the new variable explained the largest proportion of the previously unexplained variability, and so the problem of collinearity of variables was minimized. The new model was then estimated, and the process repeated. It was continued until the addition of a further variable proved to be statistically insignificant at the 5% level. This procedure is in the spirit of Atkinson's added variable method, and yielded a preliminary set of potential needs variables. To this set was added a small number of additional variables, felt *a priori* to be important determinants of health care needs, but not detected in the preliminary statistical screening.

5.14 The model arrived at using this strategy was specified as the full "unrestricted model", which was statistically well-specified at the 5% level. In the acute sector, this model comprised the supply variables together with 30 needs variables. A check was made at this stage to ensure that the supply variables were indeed endogenous (they were always found to be so).

5.15 In order to converge to a parsimonious model, an attempt was then made to exclude variables. The criteria for selecting candidates for exclusion included the statistical significance of the variable in the model, and the size of its beta coefficient - the coefficient on the variable standardized to take zero mean and unit standard deviation. A check was made to determine whether the exclusion of the chosen variable was valid under the restriction test noted above at the 0.1% significance level. In addition, to be accepted, the new restricted model was required to pass the specification test at the 0.1% significance level.

5.16 Because of the high level of correlation between many of the variables, the variables selected were to some extent arbitrary, in the sense that a different combination of variables may have been able to yield roughly similar explanatory power. We must therefore emphasize that all models selected are to some extent a

compromise between full statistical rigour and usefulness for policy purposes, and that not too much meaning should be attached to the actual variables selected, as there may be other combinations of variables which capture the same needs characteristics. That is, the variables selected are proxies for unmeasurable social factors which could be equally successfully captured by other variables, so the precise variable selected is less important than the social factor it partially captures.

5.2 Developing a resource allocation formula

5.17 In developing a resource allocation formula, we wish to correct for variations in supply between areas. Effectively, this means assuming that all supply in an area is at some national average level appropriate to the level of needs found in the area. In calculating a measure of relative need, therefore, the variation in utilization due to variation in supply variables should be considered only to the extent that supply reflects variations in legitimate need for health care. The requirement is to develop a measure of "normative utilization": the utilization that would obtain in an area if the response to its needs was at the national average level.

5.18 In order to illustrate the problem, consider the model of utilization developed in Chapter 3, and reproduced as Figure 5.1. Variations in utilization U arise because of variations in needs N and variations in supply S. Normative utilization is that part of utilization which is attributable to needs alone. Now needs can influence utilization in two ways: first through a direct impact, as indicated by the arrow *a*, and second as mediated through supply (arrows *b* and *c*). The analytic task is to find that part of the supply effect *c* which is attributable to factors X unrelated to needs indicators N, and to remove that part of the supply effect from the model. Thus we wish to take account of supply. But we only wish to consider supply to the extent that it reflects legitimate health care needs.

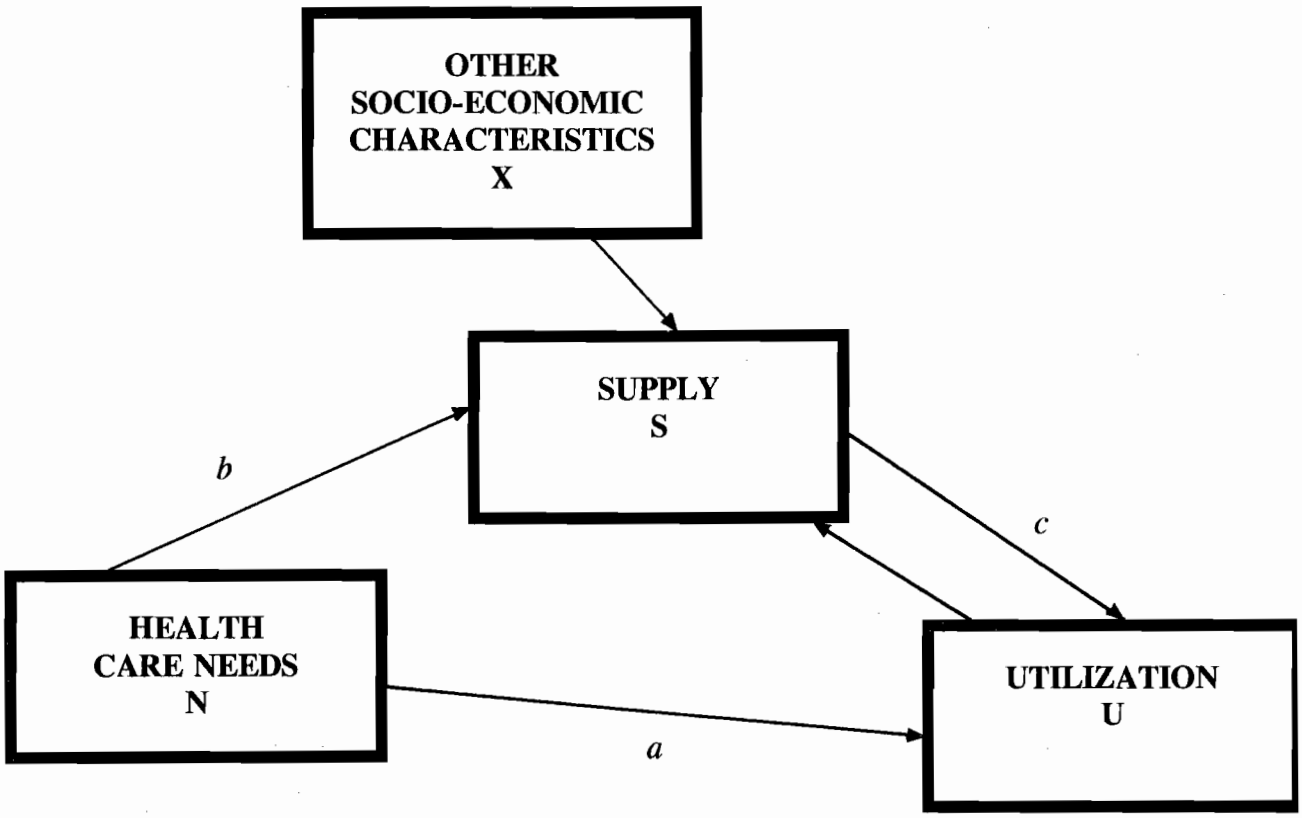


Figure 5.1: The simplified model of demand for health care

5.19 The purpose of the two stage least squares modelling exercise was to isolate legitimate needs variables N which have an unambiguous statistical relationship with utilization. However, the equations derived from this part of the study are not directly helpful from the viewpoint of developing a formula, as the coefficient on supply will reflect both legitimate needs N and extraneous variables X. A technology is therefore needed to isolate the impact of needs alone on utilization, either directly, or as mediated through supply. The remainder of this Section explains the solution to this problem adopted in this study.

5.20 Eliminating supply S from equations (5.1) and (5.2), the following equation is

derived:

$$U_i = g_1(N_i, g_2(N_i, U_i, X_i)) \tag{5.6}$$

This can be solved to yield the following expression for utilization:

$$U_i = g_3(N_i, X_i) \tag{5.7}$$

Equation (5.7) is known as the "reduced form" expression for U, which explains U in terms of legitimate needs variables N and more general socio-economic variables X. Then it is noteworthy that the impact of needs N on utilization U is given by the total derivative:

$$\frac{dU}{dN} = \frac{\partial g_3}{\partial N} + \frac{\partial g_3}{\partial X}\frac{\partial X}{\partial N} \tag{5.8}$$

That is, the total effect of needs N on utilization is found by examining both the direct effect of N on U, and any indirect effect on U associated with X, if X is correlated with N.

5.21 The purpose of the two stage least squares modelling exercise was to isolate legitimate "needs" drivers N of utilization. In order to develop a measure of normative utilization, it is necessary to estimate the response of utilization to those needs variables alone, taking account of any covariance the needs variables might have with broader socio-economic circumstances. This is achieved by undertaking an ordinary least squares regression of utilization on the needs variables identified in the modelling work described above. The coefficient on each needs variable will then capture its direct and indirect effect on utilization.

5.22 The legitimate needs variables N (in logarithmic form) are therefore entered in an

OLS regression of utilization as follows:

$$U_i = \alpha + \sum_{j=1}^{m} \beta_j N_{ij} + \epsilon_i \qquad (5.9)$$

The coefficients in the resulting equation reflect the *total* impact of needs on utilization, and can therefore be used as the basis for a resource allocation formula. In contrast, in the Review of RAWP the needs coefficients recommended for use were taken from a regression of utilization on both needs and supply. The strategy used in this study allows for the fact that variations in existing supply might already to some extent reflect variations in legitimate health care needs. By regressing utilization on needs alone, that part of needs which is correlated with supply has been taken into account.

5.3 Multilevel modelling of utilization

5.23 A fundamental assumption of conventional statistical modelling of the sort described in Section 5.1 is that the residuals, or unexplained variations from the estimated model, are independently distributed. However, it is quite possible that there exist systematic effects of administrative areas (or "levels") on utilization. For example, a DHA policy to carry out some minor procedures in outpatient clinics may tend to depress inpatient utilization rates in all wards within the DHA, or DHA practice in defining completed consultant episodes may have a systematic impact on utilization rates throughout the District. In practice, therefore, it is plausible to suggest that there may be clustering of residuals within administrative areas. Multilevel modelling techniques - originally developed for use in the education sector - explicitly take account of such clustering and are designed to search for higher level effects, such as those caused by DHA, FHSA or RHA policy, and to search for explanations for any such effects (Paterson and Goldstein, 1991). They effectively seek to distinguish between inter- and intra-

district variations, and have been used elsewhere to explore variations in health care utilization (Gatsonis *et al*, 1993). From a technical viewpoint, where clustering exists, ordinary regression methods might lead to biased estimates of standard errors and so lead to incorrect inferences (see Appendix E).

5.24 The basic structure of the multilevel model is the same as the conventional statistical model described in Section 5.2. However, within that structure and before considering the appropriate functional form of the equation, equation (5.9) is extended to the following:

$$U_{ik}=\alpha+\sum_{j=1}^{m}\beta_j N_{ijk}+\epsilon_{ik}+\nu_k \qquad \textbf{(5.10)}$$

where k represents a District (or other administrative area). Moreover, instead of assuming that there exists a single value for each of the α and β parameters to be estimated, it is assumed that the parameters might vary between administrative areas. Thus, emerging from the multilevel analysis is an estimate of the extent to which DHA, FHSA and RHA effects exist, and a description of how the parameters of the utilization equation vary between administrative areas. In effect, for each explanatory variable in the model, the multilevel analysis produces an estimate of the slope and intercept in the utilization equation specific to each administrative area.

5.25 In addition, by examining the area-specific data described in Section 4.2.5, the multilevel analysis can also seek to explain the variation in parameter estimates between areas. If significant effects are found, further analysis may indicate important policy and other area-specific influences on utilization, and may suggest variables for which ward level proxies should be sought in order to improve the ward level model of utilization.

References

Gatsonis C, Sharon-Lise N, Lin C and Morris C (1993) Geographic variation of procedure utilization, *Medical Care*, 31, YS54-59.

Paterson L and Goldstein H (1991) New statistical methods for analysing social structures: an introduction to multilevel models, *British Educational Research Journal*, 17(4), 387-393.

6. RESULTS

6.1 In deriving the recommended models, literally hundreds of alternative statistical specifications were tested. As with all statistical modelling, judgements had to be made about the merits of competing models. It should be emphasized that the necessary decisions were taken in full consultation with the study team's advisory groups, and that the models were developed in ignorance of their resource allocation consequences. In this Section we discuss in some detail the development of a model for the acute services in Section 6.1. Section 6.2 gives in less detail the equivalent models in the non-acute specialties.

6.1 An acute specialties model

6.2 The acute specialty groups are those described in Section 4.3 (that is, all acute specialties, including the newborn, but excluding maternity). The dependent utilization variable was the ratio of estimated costs to expected costs for the year 1990/91. The four supply variables, described in Section 4.2, were access to NHS hospitals (ACCNHS); access to general practitioners (ACCGPS); provision of residential and nursing homes (HOMES*); and access to private hospitals (ACCPRI). Natural logarithms were taken of all variables. The supply variables were found to be endogenous to the model, which was therefore estimated using two stage least squares. Needs variables were now added to the model, in accordance with the procedure described in Chapter 5. Dummy intercept variables for the Regional Health Authorities were found to be necessary at each stage of the modelling procedure. This process resulted in an unrestricted model of utilization containing 30 needs variables. The model was then restricted by omitting needs variables until the model shown in Table 6.1 was found (the Regional dummies are not shown).

	DF	Sum of Squares	Mean Square
Regression	26	135.14892	5.1980355
Residuals	4913	143.87539	.0292846

F = 177.50046 Signif F = .0000

---------- Variables in the Equation ----------

Variable	B	SE B	Beta	T	Sig T
Supply variables					
ACCNHS	.024387	.052158	.033886	.468	.6401
ACCGPS	.181345	.118829	.096191	1.526	.1270
HOMES*	.189693	.064177	.056507	2.956	.0031
ACCPRI	.149492	.039652	.375152	3.770	.0002
Needs variables					
DENSITY	-.027184	.006781	-.171982	-4.009	.0001
MANUAL	.089583	.013652	.140127	6.562	.0000
OLDALONE	.080273	.024673	.055947	3.254	.0011
S_CARER	.088610	.019027	.110350	4.657	.0000
UNEMP	.062195	.012855	.133375	4.838	.0000
PRIVRENT	.016626	.006584	.054345	2.525	.0116
BLACK*	.175224	.075286	.036992	2.327	.0200
SIR074	.057008	.027349	.070372	2.084	.0372
SMR074	.120520	.021336	.115851	5.649	.0000
(Constant)	4.707407	.230536		20.419	.0000

Table 6.1: Intermediate model of utilization, acute services 1990/91

The variables included were:

ACCNHS	Access to NHS acute beds;
ACCGPS	Access to general practitioners;
HOMES*	Proportion of population aged 75+ not in nursing or residential homes;
ACCPRI	Access to private hospital beds;
DENSITY	Persons divided by hectares;
MANUAL	Proportion in households with head in manual social classes;
OLDALONE	Proportion of pensionable age living alone;
S_CARER	Proportion of dependants in single carer households;
UNEMP	Proportion of economically active unemployed;
PRIVRENT	Proportion in private rented accommodation;
BLACK*	Proportion not in black ethnic groups;
SMR074	SMR for ages 0-74;
SIR074	Standardized illness ratio for ages 0-74.

Recall that, using a logarithmic model, the B coefficients can be interpreted as the elasticities of the associated variable. The model passes the specification test ($\chi^2(31) = 57.0$, $p > 0.001$) and, from the restriction test, any larger model is

statistically unnecessary. The four supply variables are indeed endogenous ($F(4,4910) = 6.7$, $p < 0.005$).

6.3 Because of the need to take the natural logarithms of all variables, those marked * are defined in the opposite sense to the usual definition. The HOMES* variable therefore reflects the proportion of those aged over 75 *not* living in homes. As a result, the coefficient on these variables should be interpreted with care. Most of the variables selected are intuitively sensible: SMR074 and SIR074 are capturing health characteristics, PRIVRENT, MANUAL and UNEMP various aspects of poverty, while OLDALONE and S_CARER appear to be capturing home circumstances. The BLACK* coefficient is less intuitively plausible, suggesting that areas with higher proportions of black residents exhibit lower utilization than expected. The study team and its advisers interpreted this finding as reflecting *supply* rather than need, perhaps because wards with large ethnic minority populations tend to be close to acute hospitals. The BLACK* variable was therefore not considered as a needs variable. The strong negative coefficients on DENSITY suggests higher utilization than expected in rural areas. As we show later, the DENSITY variable becomes insignificant at the last stage of the analysis. Again, therefore, this phenomenon was interpreted as reflecting residual supply characteristics not captured in our chosen supply variables. For example, there might be higher levels of unmet demand in urban areas than in rural areas.

6.4 Recall from Chapter 5 that the purpose of the initial two stage least squares analysis was to develop an understanding of indicators of need for NHS facilities independent of supply. While the model in Table 6.1 appears to offer a good statistical description of the determinants of utilization, the number of variables used makes it unwieldy for policy purposes. In addition, some of the variables, although statistically significant, have relatively little influence on utilization (their beta coefficients are small). We therefore sought to restrict this model to a more manageable form. Numerous variants of the model omitting some of the needs

variables were therefore tested. The most parsimonious model exhibiting reasonable statistical properties is that shown in Table 6.2.

	DF	Sum of Squares	Mean Square
Regression	23	134.69869	5.8564646
Residuals	4916	141.43557	.0287705

F = 203.55826 Signif F = .0000

Variables in the Equation

Variable	B	SE B	Beta	T	Sig T
Supply variables					
ACCNHS	.039633	.048276	.055069	.821	.4117
ACCGPS	.332842	.077140	.176550	4.315	.0000
HOMES*	.132420	.061604	.039446	2.150	.0316
ACCPRI	.111208	.035712	.279078	3.114	.0019
Needs variables					
DENSITY	-.035102	.005033	-.222074	-6.975	.0000
MANUAL	.092509	.012391	.144704	7.466	.0000
OLDALONE	.104009	.023432	.072490	4.439	.0000
S_CARER	.081852	.016490	.101934	4.964	.0000
UNEMP	.073764	.011402	.158184	6.469	.0000
SMR074	.134619	.020416	.129404	6.594	.0000
(Constant)	4.700487	.188911		24.882	.0000

Table 6.2: Parsimonious model of utilization, acute services 1990/91

6.5 Table 6.2 gives the most parsimonious model we found from which we could not exclude a variable at the 0.1% level. The model exhibits some evidence of misspecification ($\chi^2(34) = 73.7$). However, with almost 5,000 observations some misspecification is to be expected. Variables reflecting proportions of persons in each age/sex category were added to this model, but only one was found to be significant, suggesting that the age/sex standardization was satisfactory.

6.6 The study team's advisers subjected the models identified above and several variants thereof to intense scrutiny. In particular, they recommended that two refinements to the needs variables under consideration should be tested. First, a variable NO_CARER was introduced, measuring the proportion of dependants (principally elderly) with no carer in the household. Second, an alternative

version of SIR074 was tested, measuring the standardized illness ratio in households only. This variable, denoted HSIR074, excludes the effect of institutions in an area, which may distort the SIR.

6.7 The introduction of the NO_CARER variable was found to affect the model very little, perhaps because much of the variability associated with it was already captured by OLDALONE . However, augmenting the parsimonious model with HSIR074 improved the model specification considerably, as shown in Table 6.3. Furthermore, we were able to confirm that this model would have emerged if we had used HSIR074 from the start, and it was therefore retained as the preferred model in the acute sector. The specification test statistic is $\chi^2(33) = 69.7$.

	DF	Sum of Squares	Mean Square
Regression	24	135.22982	5.6345758
Residuals	4915	142.53164	.0289993

F = 194.30030 Signif F = .0000

---------- Variables in the Equation ----------

Variable	B	SE B	Beta	T	Sig T
ACCNHS	.023446	.049702	.032578	.472	.6371
ACCGPS	.376911	.080909	.199925	4.658	.0000
HOMES*	.129046	.067104	.038441	1.923	.0545
ACCPRI	.121444	.036218	.304766	3.353	.0008
DENSITY	-.036979	.005071	-.233950	-7.292	.0000
MANUAL	.073344	.013846	.114726	5.297	.0000
OLDALONE	.091489	.024015	.063764	3.810	.0001
S_CARER	.057695	.017831	.071850	3.236	.0012
UNEMP	.047465	.013362	.101786	3.552	.0004
HSIR074	.109037	.029128	.136801	3.743	.0002
SMR074	.117935	.021044	.113366	5.604	.0000

Table 6.3: Acute services model including Household Standardized Illness Ratio

6.8 With the exception of access to NHS hospitals, the supply variables are important. Increased accessibility to private hospitals and GPs is positively associated with use of NHS inpatient facilities. The effect of nursing and residential homes is less strong, but suggests that higher provision is associated with depressed utilization (recall that HOMES* is defined as proportion *not* in homes). All the needs

indicators indicate strong, intuitively plausible effects on utilization.

6.9 Thus the two stage least squares procedure identified what in our judgement is a good model of the determinants of variations in utilization, and an understanding of the most suitable indicators of the need for health care. The next stage of the analysis was to use the results to develop a formula which is sensitive not only to legitimate health care needs, but also to supply to the extent that it reflects those needs. As discussed in Section 5.2, this was achieved by undertaking an OLS regression of utilization on the "legitimate" needs variables identified in the two stage least squares analysis.

6.10 The coefficients on DENSITY and MANUAL were found at the OLS stage to be statistically insignificant, and so these variables were therefore removed from the model. The OLS regression of utilization on the remaining five needs indicators is shown in Table 6.4. For this part of the analysis, the study team had available both 1990/91 and 1991/92 HES data, and so the OLS analysis was undertaken on both years' data combined. Again, regional dummies are excluded from the Table.

```
Multiple R            .73558
R Square              .54108
Adjusted R Square     .53941
Standard Error        .14354

Analysis of Variance
                    DF        Sum of Squares      Mean Square
Regression          18             119.87234          6.65957
Residual          4934             101.66886           .02060

F =     323.21599       Signif F =  .0000

------------------ Variables in the Equation ------------------

Variable                B         SE B        Beta           T  Sig T

UNEMP             .050448      .010221     .120200       4.936  .0000
OLDALONE          .132275      .016016     .102643       8.259  .0000
SMR074            .138722      .016612     .148241       8.351  .0000
S_CARER           .043181      .014341     .059761       3.011  .0026
HSIR074           .202276      .020309     .282048       9.960  .0000
```

Table 6.4: Regression of utilization on needs, acute specialties, 1990/92

6.11 The coefficients on most of the needs variables in the acute model in Table 6.4 are higher than those reported for the model in Table 6.3. Examination of the beta coefficients suggests that OLDALONE, SMR074 and HSIR074 are the most important determinants of utilization. The R^2 statistic indicates that the chosen acute sector needs variables account for 54% of the variance in age/sex adjusted utilization, but it should be noted that at this stage we are not necessarily seeking to develop a model which offers the best fit to the data. This could be achieved simply by indiscriminately adding further socio-economic variables. Instead, we have tried to identify a model that explains that part of the variation in utilization caused only by legitimate needs.

6.12 The final stage in the analysis was to undertake a multilevel analysis, to determine whether there are systematic DHA effects on utilization. The multilevel modelling approach serves two purposes: it provides unbiased estimates of the coefficients in the presence of significant district clustering; and it facilitates the search for variables which might account for inter-district variation. Whilst the latter is of considerable interest in attempting to understand the causal processes underlying

health service utilization, for the purposes of producing a formula for national resource allocation, the former is our main concern.

6.13 The first point to establish is whether there is variation attributable to the district level. This is confirmed by noting that, when the variance between observational units is partitioned between the synthetic ward level and the DHA district level, the estimate of the variance component at the district level is 9.1 times its standard error. Across all Districts, ignoring the variation between population sizes, about 44% of the variance in the utilization rate is attributable to inter-district variation. It is therefore important to re-estimate the OLS equation of Table 6.4 using multilevel estimation methods.

6.14 The multilevel results are presented in column (a) of Table 6.5. Effectively, they reproduce the model estimated for Table 6.4, but allowing for inter-district variation in the coefficients. It can be seen that the results are broadly similar to the OLS results, which are reproduced as column (b). For the reasons discussed in Section 5.3, we believe that the multilevel results represent the best estimates of the national average relationship between the chosen needs variables and utilization.

	ML Model (a)	OLS (b)
Constant	-1.651 (0.1168)	-1.188
SMR074	0.1619 (0.0131)	0.139
HSIR074	0.2528 (0.0183)	0.202
OLDALONE	0.0765 (0.0130)	0.132
S_CARER	0.0436 (0.0121)	0.043
UNEMP	0.0287 (0.0092)	0.050

Table 6.5: Acute multilevel model 1990/92 (standard errors in parentheses)

6.2 The non-acute specialties models

6.15 As discussed in Chapter 4, the long stay specialties - mental handicap, mental illness and geriatrics - posed particular problems because of the difficulty of developing a satisfactory dependent variable. In the years being studied, many long stay patients were being discharged into the community, and so the pattern of utilization this gave rise to may not be an accurate reflection of need. The models developed in the non-acute sector were therefore viewed with some caution. Nevertheless, we believe that in the absence of alternative evidence they offer the best available indication of need in the non-acute sector.

6.16 Numerous two stage least squares models were developed, for the following dependent variables:

mental handicap;
mental illness;
geriatrics;
mental illness and geriatrics combined;
all three combined.

The models were developed using both standard costs and estimated costs, and for

both 1990/91 and 1991/92. After close scrutiny, the study team and its advisers chose to focus on standard costs (for the reasons discussed in Chapter 4) and the year 1991/92 (for which the data were thought to be more reliable). Furthermore, it was decided that the study methodology was inappropriate for the mental handicap specialty. There was considerable clustering in this specialty, with about one third of wards exhibiting no utilization. We were therefore unable to develop a model for mental handicap, and suggest that further work is undertaken in this area using alternative methodologies.

6.17 Access to non-acute NHS beds (ACCNHSN) replaced access to acute beds as the measure of NHS inpatient supply. Again, logarithms were taken of all variables. Adequate models were developed for both mental illness and geriatrics. The results of the two stage least squares analysis for mental illness are shown in Table 6.6.

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	25	510.5774	20.423096
Residual	4905	1234.8052	.251744

F = 81.11858 Signif F = .0000

Variables in the Equation

Variable	B	SE B	Beta	T	Sig T
ACCNHSN	-.332903	.128020	-.174743	-2.600	.0093
ACCGPS	.322967	.080632	.135998	4.005	.0001
HOMES*	-.767352	.167465	-.091071	-4.582	.0000
ACCPRI	.158216	.101552	.158060	1.558	.1193
LONE_P	.145440	.025642	.136344	5.672	.0000
NO_CARER	.118899	.032907	.070155	3.613	.0003
NEW_COM	.058307	.012042	.113052	4.842	.0000
OLDALONE	.204529	.074654	.056818	2.740	.0062
SMR074	.253698	.060837	.097192	4.170	.0000
P_SICK	.212669	.031778	.178958	6.690	.0000
PCTURB	-.109199	.040727	-.042059	-2.681	.0074
STUDENTS	-.069452	.032908	-.038960	-2.111	.0349
(Constant)	1.652328	.549250		3.008	.0026

Table 6.6: Mental illness model, 1991/92

The variables included in the mental illness model were as follows:

ACCNHSN	Access to NHS non-acute beds;
ACCGPS	Access to general practitioners;
HOMES*	Proportion of population aged 75+ not in nursing or residential homes;
ACCPRI	Access to private hospital beds;
LONE_P	Proportion in households headed by a lone parent;
NO_CARER	Proportion of dependants with no carer
NEW_COM	Proportion in persons born in New Commonwealth;
OLDALONE	Proportion of pensionable age living alone
SMR074	Standardized mortality ratio (SMR) for ages 0-74;
P_SICK	Proportion of adult population permanently sick;
PCTURB	Percentage of population living in "urban" enumeration districts (as defined by Department of Environment);
STUDENTS	Proportion of 17 year olds who are students.

6.18 The supply variables are confirmed as endogenous ($F(4,4910) = 6.4$, $p<0.005$). There is some evidence of misspecification ($\chi^2(32) = 67.5$), but this must be expected with so many observations. Again, age/sex standardization appears to have been satisfactory. In interpreting all models for long stay specialties it is important to bear in mind the warning about long stay specialties discussed in paragraph 6.15. The HES database used to develop the model in Table 6.6 records discharges in 1991/92, many of which related to episodes of very long duration. The apparent pattern of need this gives rise to may therefore be misleading, particularly if discharge destinations - the basis for HES ward of residence - are clustered into certain wards.

6.19 Table 6.7 presents the results of regressing mental illness utilization in years 1990/91 and 1991/92 combined on the needs variables identified in Table 6.6, omitting PCTURB and STUDENTS, the coefficients of which became insignificant at this stage. Again, note the relatively high value of R^2.

```
Multiple R            .63378
R Square              .40168
Adjusted R Square     .39938
Standard Error        .38448

Analysis of Variance
                    DF        Sum of Squares      Mean Square
Regression          19             489.57810         25.76727
Residual          4933             729.24734           .14782

F =     174.31173        Signif F =  .0000

------------------ Variables in the Equation ------------------

Variable                B         SE B        Beta           T  Sig T

LONE_P            .152412      .016672     .169975       9.142  .0000
NO_CARER          .113106      .023946     .079438       4.723  .0000
NEW_COM           .048625      .006536     .112233       7.439  .0000
OLDALONE          .375256      .054216     .124145       6.922  .0000
SMR074            .226731      .044695     .103296       5.073  .0000
P_SICK            .267122      .023300     .267284      11.465  .0000
(Constant)        .730580      .271406                   2.692  .0071
```

Table 6.7: Regression of utilization on needs, mental illness, 1990/92

6.20 When the multilevel analysis was undertaken, there was once again a substantial district effect, with the variance component at DHA district level being 9.0 times its standard error; and the proportion of overall variance attributable to the district level being approximately 38%. It was therefore important to re-estimate the OLS equation of Table 6.7 using multilevel estimation methods.

6.21 The multilevel results are presented in Table 6.8. Column (a) presents the model estimated for Table 6.7, but allowing for inter-district variation. The results suggest a similar response to needs to that indicated in Table 6.7 (reproduced as column (b)). The multilevel model is recommended as a basis for a national formula in the mental illness specialty.

	ML - Model (a)	OLS (b)
Constant	1.1270 (0.2414)	0.7306
LONE_P	0.1846 (0.0149)	0.1524
NO_CARER	0.1431 (0.0213)	0.1131
NEW_COM	0.1073 (0.0070)	0.0486
OLDALONE	0.3609 (0.0468)	0.3753
SMR074	0.2426 (0.0382)	0.2267
P_SICK	0.2616 (0.0215)	0.2671

Table 6.8: Mental illness multilevel model 1990/92 (standard errors in parentheses)

6.22 It was also possible to develop satisfactory models for the geriatric specialty and for mental illness and geriatrics specialties combined. The multilevel and OLS models for 1990/92 (after deletion of statistically insignificant variables) are shown in Tables 6.9 and 6.10. The explanatory power of the OLS combined model is good (R^2 = 0.47), while that of the geriatrics model is relatively poor (R^2 = 0.17). The new variables used in these two models are:

LONE_P2*	Proportion of families which are *not* lone parent families;
NON_SC	Proportion in households in non-self-contained accommodation;
NEW_COM2	Proportion in households with head born in New Commonwealth;
SMR75+	Standardized mortality ratio (SMR) for ages 75+.

	ML - Model (a)	OLS (b)
Constant	0.3550 (0.1628)	0.4961 (0.0995)
MANUAL	0.2501 (0.0236)	0.3646 (0.0352)
LONE_P2*	-1.368 (0.155)	-1.1986 (0.2221)
NEW_COM	0.0790 (0.0097)	0.0825 (0.0130)

Table 6.9: Geriatrics multilevel and OLS models 1990/92 (standard errors in parentheses)

	ML - Model (a)	OLS (b)
Constant	-0.0599 (0.1943)	-0.5010 (0.2198)
NON_SC	-2.3540 (0.0472)	-2.5562 (0.4467)
NEW_COM2	0.0873 (0.0054)	0.0550 (0.0052)
LONE_P	0.1802 (0.0118)	0.1531 (0.0132)
NO_CARER	0.1359 (0.0148)	0.0972 (0.0171)
P_SICK	0.1570 (0.0175)	0.1239 (0.0190)
SMR074	0.2202 (0.0308)	0.2404 (0.0366)
SMR75+	0.1169 (0.0231)	0.1167 (0.0283)
PRIVRENT	-0.0202 (0.0072)	-0.0178 (0.0079)
MANUAL	0.0374 (0.0222)	0.1149 (0.0237)

Table 6.10: Table 6.9: Geriatrics and mental illness combined multilevel and OLS models 1990/92 (standard errors in parentheses)

6.3 Sensitivity analysis

6.23 The models identified in Sections 6.1 and 6.2 are the result of a thorough statistical analysis of the determinants of one measure of utilization. In the course of developing those models, as we have explained, many judgements had to be made. It was therefore important for the study team to test the sensitivity of the results to assumptions made during the modelling process. As discussed, numerous models were developed in the course of the study, and the team's advisers played an active role in scrutinizing these variants and recommending alternative specifications. Thus the principle of testing the sensitivity of the models to different assumptions was intrinsic to the conduct of the study.

6.24 In addition, a more formal sensitivity analysis was undertaken, in which the robustness of the acute model with respect to different assumptions was examined in some detail. The model chosen for such scrutiny was the OLS version of the preferred model described in Table 6.4, which was re-estimated using different subsets of the data for 1990/91. The consequences for the size and statistical significance of the coefficients of the model are described in this Section. In addition, it was important to test the robustness of the results to the chosen measure of utilization. We therefore examined the effects of using episodes, bed days and standard costs (rather than estimated costs) as the dependent variable. All Tables are shown at the end of this Chapter.

6.25 For each alternative model, the coefficients and standard errors are reported in conjunction with the R^2. In summary, six issues relating to sensitivity analysis are reported:

(a) variations between Regions;
(b) variations between specialty groups;
(c) restricting the analysis to elective admissions;

(d) restricting the analysis to over-65 age groups;

(e) variations between "high" and "low" needs areas;

(f) substituting different measures of resource use for estimated costs.

These are now considered in turn.

6.26 The OLS model was run Region by Region (using the 14 Regions as defined before the restructuring of April 1994). The results are reported in Table 6.11. It should be noted that the two stage least squares models (not reported here) remain well specified in most regions, suggesting that the structure of the model is robust between regions. Table 6.11 suggests that the coefficients of the OLS models are unstable. However, the needs variables are highly collinear, so much of the apparent instability is caused by the burden of the model arbitrarily shifting from one variable to another. The model shows particularly high R^2 values in North Western and North East Thames. This may be because these RHAs are responding sensitively to the chosen needs factors, or because supply is particularly closely linked to needs in those areas. The model offers particularly low explanations of utilization in Oxford and South Western.

6.27 Table 6.12 reports the model as applied to the four former Thames Regions in aggregate. For estimated costs, the coefficient on lone carer becomes negative and the coefficients on SMR and unemployment are higher than for the national model. As might be expected, the standard costs model suggests less sensitivity to needs than the estimated costs model.

6.28 Table 6.13 reports the consequences of applying the model to the two principal acute specialty groups: surgical (specialty group 1) and medical (specialty group 3). The models are reasonably consistent with the general acute model. In general, more weight is attached to the needs variables in the medical model. These results are consistent with received epidemiological wisdom.

6.29 Table 6.14 examines the consequences of restricting the model to elective admissions. The second line of the Table examines the model restricted to a subset of elective admissions: those for which GP fundholding procedures were undertaken.

6.30 Restricting the model to those aged 65 and over yields the results shown in Table 6.15. The model explains only 22% of variance, and suggests lower response to needs than the general model.

6.31 We explored the possibility of non-linearities in the relationship between needs variables and the demand for health care by splitting the dataset into two on the basis of the ward's SMR074. Table 6.16 shows the results for the model restricted first to the "high needs" sample (SMR > 100) and then to the "low needs" sample (SMR < 100). The importance of SMR074 and lone carers is strongest amongst high needs areas, while OLDALONE and SIR074 are most important amongst low needs areas. This analysis offers some evidence of different determinants of demand amongst areas with different needs characteristics. Further work might explore this phenomenon in more depth.

6.32 An important component of the sensitivity analysis was to test the consequences of using alternative measures of utilization. Table 6.17 therefore reports the implications of substituting episodes, bed days and standard costs as the dependent variable. Using episodes, the SIR assumes increased importance, while with bed days it is the other needs variables that exhibit increased coefficients. Estimated costs are effectively a combination of episodes and bed days, weighted in accordance with the costs shown in Table 4.5 (page 66). Use of episodes or bed days gives rise to predictions of utilization which are similar to those produced by the estimated costs model, suggesting that the model of needs is reasonably robust to the choice of cost weightings.

6.33 The standard costs of each episode were calculated using East Cheshire cost data, as explained in Chapter 4 (page 64). Using this dependent variable, all the needs variables with the exception of SIR074 assume less importance than in the estimated costs model, and the standard costs model is therefore less sensitive to needs. The argument for using estimated rather than standard costs is that variations in length of stay (and therefore costs) might be the result of legitimate social needs variations rather than variations in clinical practice and efficiency. These results do indeed suggest that length of stay is influenced by needs considerations.

6.4 Conclusions

6.34 We have been able to draw on a much more extensive data set, more appropriate statistical methodology, and sounder theoretical understanding of the processes involved than our predecessors, and are convinced that this is the best that can be done with the current data available at this level of analysis. Although in developing these models we have had to make numerous judgements, we have sought to do this in full consultation with our advisers, and with no advance knowledge of the resource allocation consequences of our choices. We have identified what appears to be a convincing model in the acute sector, which is consistent with available epidemiological evidence. Modelling non-acute utilization was always likely to be more problematic, but we believe that the models reported here are a marked advance on the current use of the square root of under-75 SMR.

Region	SMR074	Elderly alone	LoneCarer	Unemploy	SIR074	R-squared
Northern	0.178 (0.078)	0.179 (0.075)	0.046 (0.067)	0.044 (0.043)	0.226 (0.080)	0.414
Yorkshire	0.275 (0.104)	-0.061 (0.097)	0.217 (0.075)	-0.022 (0.047)	0.132 (0.113)	0.279
Trent	-0.012 (0.067)	0.380 (0.062)	0.120 (0.058)	0.125 (0.042)	0.065 (0.063)	0.454
East Anglia	0.204 (0.078)	0.209 (0.073)	0.104 (0.073)	-0.028 (0.057)	0.288 (0.124)	0.362
NW Thames	0.199 (0.069)	0.081 (0.064)	0.042 (0.063)	0.161 (0.054)	0.168 (0.098)	0.388
NE Thames	0.150 (0.066)	0.282 (0.061)	-0.055 (0.048)	0.150 (0.045)	0.173 (0.094)	0.524
SE Thames	0.435 (0.059)	-0.008 (0.059)	0.066 (0.051)	-0.004 (0.049)	0.092 (0.083)	0.417
SW Thames	0.033 (0.100)	0.354 (0.088)	0.099 (0.097)	0.195 (0.079)	-0.012 (0.138)	0.246
Wessex	0.202 (0.074)	-0.015 (0.068)	0.132 (0.071)	-0.015 (0.061)	0.358 (0.106)	0.333
Oxford	0.062 (0.091)	0.062 (0.091)	-0.003 (0.079)	0.192 (0.075)	-0.004 (0.131)	0.141
S Western	-0.026 (0.067)	0.165 (0.075)	0.027 (0.076)	0.127 (0.053)	-0.049 (0.097)	0.106
West Midlands	0.185 (0.057)	0.120 (0.057)	0.175 (0.047)	0.041 (0.027)	0.054 (0.058)	0.449
Mersey	0.051 (0.089)	0.080 (0.091)	0.016 (0.134)	0.116 (0.069)	0.239 (0.122)	0.479
N Western	0.078 (0.060)	0.027 (0.065)	0.048 (0.047)	0.087 (0.027)	0.225 (0.059)	0.549
All regions (no dummy variables)	0.139 (0.017)	0.132 (0.016)	0.043 (0.014)	0.050 (0.010)	0.202 (0.020)	0.541
All regions (with dummies)	0.141 (0.020)	0.149 (0.019)	0.085 (0.017)	0.074 (0.012)	0.124 (0.024)	0.481

Table 6.11: Regional analysis - acute model

Dependent variable	SMR074	Elderly alone	LoneCarer	Unemploy	SIR074	R-squared
Estimated costs	0.211 (0.038)	0.160 (0.036)	-0.076 (0.032)	0.189 (0.028)	0.123 (0.053)	0.370
Standard costs	0.177 (0.034)	0.099 (0.062)	-0.070 (0.030)	0.087 (0.026)	0.167 (0.048)	0.250
All regions (estimated costs)	0.141 (0.020)	0.149 (0.019)	0.085 (0.017)	0.074 (0.012)	0.124 (0.024)	0.481

Table 6.12: Thames regions - acute model

Speciality group	SMR074	Elderly alone	LoneCarer	Unemploy	SIR074	R-squared
Surgical	0.171 (0.022)	0.058 (0.021)	0.074 (0.019)	0.012 (0.013)	0.162 (0.026)	0.411
Medical	0.159 (0.030)	0.288 (0.029)	-0.003 (0.026)	0.146 (0.018)	0.257 (0.036)	0.461
All groups	0.141 (0.020)	0.149 (0.019)	0.085 (0.017)	0.074 (0.012)	0.124 (0.024)	0.481

Table 6.13: Speciality group - (estimated costs)

Dependent variable	SMR074	Elderly alone	LoneCarer	Unemploy	SIR074	R-squared
Elective procedures	0.088 (0.021)	-0.006 (0.020)	0.134 (0.018)	-0.015 (0.013)	0.165 (0.025)	0.389
Fund-holding procedures	0.109 (0.034)	-0.169 (0.033)	-0.119 (0.029)	-0.011 (0.021)	0.474 (0.041)	0.405
All acute	0.141 (0.020)	0.149 (0.019)	0.085 (0.017)	0.074 (0.012)	0.124 (0.024)	0.481

Table 6.14: Elective procedures - acute model

Dependent variable	SMR074	Elderly alone	LoneCarer	Unemploy	SIR074	R-squared
Elderly costs	0.113 (0.030)	0.223 (0.029)	-0.028 (0.026)	0.072 (0.018)	0.049 (0.036)	0.221
All costs	0.141 (0.020)	0.149 (0.019)	0.085 (0.017)	0.074 (0.012)	0.124 (0.024)	0.481

Table 6.15: Elderly costs - acute model

Dependent variable	SMR074	Elderly alone	LoneCarer	Unemploy	SIR074	R-squared
SMR > 100	0.209 (0.036)	0.081 (0.026)	0.127 (0.023)	0.080 (0.016)	0.078 (0.032)	0.398
SMR < 100	0.113 (0.031)	0.194 (0.027)	0.028 (0.025)	0.082 (0.019)	0.165 (0.035)	0.283
All cases	0.141 (0.020)	0.149 (0.019)	0.085 (0.017)	0.074 (0.012)	0.124 (0.024)	0.481

Table 6.16: Acute model with observations split between wards with low and high SMR

Dependent variable	SMR074	Elderly alone	LoneCarer	Unemploy	SIR074	R-squared
Episodes	0.100 (0.018)	0.028 (0.017)	0.098 (0.015)	0.040 (0.011)	0.203 (0.021)	0.504
Bed days	0.152 (0.032)	0.159 (0.031)	0.107 (0.028)	0.119 (0.020)	0.083 (0.040)	0.322
Standard costs	0.096 (0.018)	0.076 (0.017)	0.064 (0.015)	0.054 (0.011)	0.133 (0.022)	0.429
Estimated costs	0.141 (0.020)	0.149 (0.019)	0.085 (0.017)	0.074 (0.012)	0.124 (0.024)	0.481

Table 6.17: Acute model with standard rather than estimated costs as the dependent variable

Note for Tables 6.11-6.17: Figures in parentheses are estimated standard errors

7. CONCLUSIONS

7.1 The purpose of this study was to seek to provide a sound empirical basis for a formula to distribute hospital and community health services resources between health authorities in England, taking into account many of the criticisms that have been levelled at previous work. Our remit was to carry out a small area analysis of hospital admissions using the most appropriate statistical analyses. In this final section, we summarize the advances made in this study, our general conclusions about the form of our preferred model, and our recommendations.

7.2 This study marks an advance on previous empirical work in three main areas: the size and quality of the dataset; the statistical techniques used; and the interpretation of results. We therefore believe that the use of the study's findings would lead to a more equitable allocation of resources in the HCHS than hitherto.

7.3 The dataset was more comprehensive than any available to previous researchers. It covered virtually the whole of England. It incorporated the 1991 Census of Population, which - as well as being contemporaneous with the utilization data - incorporated for the first time a question on limiting long standing illness. There were available a larger number of supply-related data items than hitherto. And it was possible to incorporate into the measures of utilization an estimate of the cost of each episode, based on its duration and specialty group.

7.4 The study team has based its work on an explicit model of demand for health care, and was able to use the statistical techniques most appropriate to estimating such a model. In particular, the use of methods such as two stage least squares regression techniques is essential if - as seems plausible - there is a feedback from utilization to supply. In addition, because many supply considerations are defined only at the level of administrative area, there is every reason to suppose that multilevel modelling techniques are necessary to capture DHA and FHSA level effects. The use of these techniques in this study offers the prospect of better

specified statistical models.

7.5 Having developed what we consider to be satisfactory statistical models of demand for health care, incorporating both supply and needs considerations, the final stage of the study was to model the link between utilization and the needs variables only. This approach differs from previous work. However, it is the only practical means of capturing the full impact of needs on utilization, some of which might already be reflected in supply. We believe that this procedure marks a conceptual advance on previous work.

7.6 In Section 2.3 we noted eight shortcomings of previous work on resource allocation. We have sought to address many of these issues. In particular:

(1) we have tested the impact of a wide range of health needs variables on utilization;

(2) we have used estimates of the resource costs of utilization specific to each episode;

(3) we have explicitly incorporated four aspects of the supply of health care into the models;

(4) using the multilevel modelling techniques, we have taken into account differences in policies and practices between health authorities;

(5) we have modelled a wide range of social circumstances as possible additional determinants of need for health care;

(6) we have sought to use technically appropriate statistical methods;

(7) we have tested the robustness of models by subjecting them to extensive sensitivity analysis.

7.7 A major criticism of previous methods that we have not been able to address is the assumption implicit in our work that the existing national allocation of resources between care groups is appropriate. However, the disaggregation into acute and non-acute sectors offers policy makers some control over the weight given to each sector.

7.8 Moreover, no study using a methodology based on utilization can capture variations in health care needs that are not reflected in utilization. Despite adjusting for supply considerations and using the appropriate statistical procedures, our methodology is vulnerable to the possibility that, for a whole range of reasons, health care needs may not be captured in hospital inpatient utilization.

7.9 We have not been able to investigate the lag structure of the relationships because of lack of data. And time limitations mean that we have not been able to explore fully the sensitivity of the models to the virtually countless possible alternative specifications. We were nevertheless able to test a wide range of models, and, whenever key judgements were required, the necessary decisions were taken in consultation with the Department of Health and its advisers. We believe that the recommended models satisfy the criteria of statistical rigour, intuitive plausibility and practicality.

7.10 The models developed in this study have indicated that, as the original RAWP report and many others suggested, supply considerations are of crucial importance in determining utilization. The strong feedback observed between supply and demand confirms the necessity for sensitive statistical treatment of the sort employed in this study. In addition, we have confirmed that some of the measures

of health status conventionally used in modelling utilization do indeed have an important role in indicating demand for health care. Moreover, we have shown that broader social factors must be taken into account when modelling utilization. The fact that supply, health status and social factors all have a clear role to play in forming demand suggests that, in making allocations to care groups, all three sets of factors should be taken into account by purchasing authorities.

7.11 The analysis has yielded national statistical average equations. The sensitivity analysis suggests that the national model is not always sustained at lower levels of aggregation. Yet, although the model changes between care groups, geographical areas and other levels of aggregation, this does not invalidate its use as the basis for a *national* allocation mechanism. The purpose of a formula is to develop a set of allocation rules which smooths out the effects of local variation in policy and practice. In other words, it should be based only on *systematic* differences in responses to needs. When using a formula derived from national data, the implicit judgement is that the national average relationship between needs variables and utilization should form the basis for national allocation of funds.

7.12 Thus the models of utilization identified in this report are based on the best available data and sound statistical techniques. We therefore believe that they constitute a fair basis for allocations to Regions for inpatient funding, and *we recommend that they should be used as the basis for national allocations to Regions.*

7.13 The study has not been able to look explicitly at the determinants of use of outpatient and community services. Clearly, in the absence of more direct measures of use, some of the data generated by the study may be of use in formulating resource allocation formulae in those sectors. In addition, we were unable to develop a formula for mental handicap. *We recommend that the relevance of the data as a basis for allocations in outpatients, community and*

mental handicap sectors should be explored further.

7.14 Also, the study has not addressed issues relating to geographical variations in the costs of providing services. It will be necessary for the allocations based on this study to be adjusted for variations in costs caused by factors such as infrastructure costs, teaching and research responsibilities, capital charges and wage costs.

7.15 After this study was commissioned, the Government announced its intention to reform the regional structure of the NHS. It is therefore possible that any formula developed might be used to make allocations directly to Districts. If the national formula is applied directly to Districts, significant divergences from current DHA allocations are likely to be observed. This result would arise with *any* realistic national formula. Nevertheless, the allocations based on this study's results will indicate a District's needs within a consistent framework. We therefore *recommend that the national formulae could be used as a basis for District targets.*

7.16 However, we also believe that no single national formula is likely satisfactorily to capture all the subtleties of variations in needs between a large number of Districts. We therefore further *recommend that, where necessary, a system for adjusting allocations to Districts in the light of local circumstances is retained.*

7.17 We believe that this study has produced the best available model of utilization, given the data and statistical resources available to the study team. There will always be variations between geographical areas caused by policy and practice which cannot be captured in a statistical model. Therefore, the small area unit of analysis will always appear to suggest that there is no single model of utilization. It is of course possible to suggest that even smaller units of analysis, such as enumeration districts, might overcome the ecological fallacy, and yield satisfactory models of demand.

7.18 However, we believe that, if an empirical basis is sought for identifying the determinants of utilization, the only method likely to yield significantly more robust and credible results than the present study is the use of long term cohort studies of individuals, which can relate the individual's needs profile to their actual use of health service resources. This issue is discussed in more detail in Appendix F. Clearly such studies are expensive. However, the sums of money redirected by the revenue allocation formulae are enormous, and it seems imperative in the interests of efficiency and equity that the allocations are well-informed. Moreover, such cohort studies would yield numerous other benefits related to health care provision. Therefore, although the current study offers a good basis for a national formula, *we recommend that serious consideration should be given to establishing a major national cohort study of health care use.*

7.19 The study has generated an invaluable dataset, which should be of interest to policy makers and researchers. *We recommend that the ward level dataset we have constructed should be released for general use as soon as possible.*

APPENDIX A: THE CREATION OF SYNTHETIC SMALL AREAS

(Prepared by OPCS in consultation with DoH)

1. The view of the Technical Group was that areas with populations averaging 7,000 were ideal for NHS Weighted Capitation analyses. Larger areas could lead to biases due to the ecological fallacy, while smaller areas would have too few events such as births, deaths, and hospital episodes for reliable estimates. Although wards were generally appropriate, the average ward size was around 5,000 with numerous smaller wards and some larger than 10,000. It was decided that wards with populations smaller than 5,000 should be grouped together. These were termed Synthetic Small Areas (SSAs).

2. An analysis was undertaken of ward level population data from the 1991 Census of Population. A frequency distribution of total population size revealed that because there were many rural wards, the average was around 5,000. A decision was taken that any wards less than 5,000 would be combined with their nearest neighbour within the same county district. This rule was likely to lead to adjacent wards being amalgamated, and a check was made by DoH on a small sample of wards to confirm that this was happening. An additional condition (eg requiring the neighbour to have a similar proportion of manual workers) would have increased the risk of non-adjacent wards being amalgamated.

3. A file of enumeration district (ED) population was obtained from OPCS Census Branch (some 110,000 EDs). These ED figures were exact figures (not Barnardised). A file of ED centroids was also obtained. The centroids used were population centred ED centroids that were estimated at the time when the census ED maps were being drawn up - not the set that was mechanically produced later by digitising software, which is not population centred. Grid references were converted into 100km distances from the square in which the Isles of Scilly

resides, and the population and centroid files were merged.

4. Population weighted centroids of electoral wards were calculated from the ED population and centroid data. Ward populations were calculated from ED populations for use in stage 5 below. The method is by weighting the ED centroids by the population and forming a weighted average of their coordinates.

5. Wards with a population of less than 5,000 were combined with their nearest neighbours within the same county district, where distance was defined as crow-fly distance. The output consisted of look-up table of census ward to final SSA destination, populations of the new SSAs, and centroids of the new SSAs.

APPENDIX B: CENSUS VARIABLES USED IN STUDY

		Table	Definition
1. Tenure			
1.1	Proportion of persons in permanent buildings owner occupied.	20	(412+413)/411
1.2	Proportions of persons in private rented	20	(414+415)/411
2. Amenities			
2.1	Proportion in households lacking bath/shower & inside WC	49	209/170
2.2	Proportion in households lacking central heating	49	222/170
2.3	Proportion in households in non-self-contained accommodation	49	235/170
3. Car ownership			
3.1	Proportion in households with no car	49	248/170
4. Overcrowding			
4.1	Proportion in households in crowded accommodation (> 1 per room)	49	(183+196)/170
5. Ethnic origin			
5.1	Proportion in households with head born in New Commonwealth	49	181/170
5.2	Proportion in non-white ethnic groups	06	{1} - 2/1
5.3	Proportion born in New Commonwealth	07	55/1
5.4	Proportion in Black ethnic groups	06	(3+4+5)/1
5.5	Proportion in Indian, Pakistani and Bangladeshi groups	06	(6+7+8)/1

6. Elderly living alone

6.1	Proportion of those aged 75+ living alone	47	(29+43+71+85)/(197+211)
6.2	Proportion of those of pensionable age living alone	47	(15+29+43+57+71+85)/169

7. Dependants

7.1	Proportion of dependants in single carer households	30	(30+70)/10
7.2	Proportion of persons in lone parent households	32	(22+23+34+35)/(10+11)
7.3	Proportion of children in lone parent households	32	(22+34)/10
7.4	Proportion of families which are economically inactive lone parent	86	280/13
7.5	Proportion of families which are lone parent with dependent child(ren)	89	12/1
7.6	Proportion of children in non-earning lone parent households	36	(12+18)/66
7.7	Proportion of children in non-earning households	36	(12+18+30+36+48)/66
7.8	Proportion of dependants with no carer	30	20/10

8. Permanently sick

8.1	Proportion of residents of working age permanently sick	08	(210-491-492-493-756-757-758-759)/ (1-282-283-284-547-548-549-550)
8.2	Proportion of adult population permanently sick	08	210/1
8.3	Age standardized permanently sick ratio (SSR)	08	Indirect, based on working age groups

9. Students

9.1	Proportion of 17 year olds who are students	08	193/3
9.2	Proportion of working age population who are students	08	(191-472-473-474-737-738-739-740)/ (1-282-283-284-547-548-549-550)

10. Migrants

10.1	Proportion of residents moving from outside l.a. district in last year	15	(1-4-5-6-7)/total population
10.2	Proportion of residents with different address to one year ago	15	1/total population

11. Limiting long-term illness

11.1	Proportion of total population with limiting long term illness	12,13	(12:1+13:3+13:4+13:7+13.8)/total pop
11.2	Age standardized illness ratio (SIR): total population (3 age ranges)	12,13	Indirect, based on all age groups
11.3	Age standardized illness ratio (HSIR): population in households (3 age ranges)	12	Indirect, based on all age groups

12. Unemployment

12.1	Proportion of economically active unemployed	08	134/20

13. Education

13.1	Proportion of persons aged 18+ with some qualification	84	4/1

14. Social class

14.1	Proportion of persons in households with head in class 1 or 2	90	(7+12)/2
14.2	Proportion of persons in households with head in manual classes	90	(22+27+32)/2
14.3	Proportion of economically active in managerial/professional SEG	92	(9+10+17+18+25+26+33+34+41+42+49+50+57+58+129+130+137+138)/(1+2)
14.4	Proportion of economically active in manual SEG	92	(81+82+89+90+97+98+105+106+113+114+121+122+145+146)/(1+2)
14.5	Proportion of economically active in non-manual SEG	92	(9+10+17+18+25+26+33+34+41+42+49+50+57+58+65+66+73+74+129+130+137+138)/(1+2)

15. Concealed families

15.1	Proportion of families that are "concealed"	88	113/105

16. Sparsity

16.1	Ratio of persons to area	01	64/Hectares
16.2	Proportion of persons in "rural" enumeration districts	-	Provided by OPCS

APPENDIX C: MEASURING ACCESSIBILITY

A fundamental need in this study was to develop a measure of the *perceived availability* of NHS inpatient services to a particular ward. This measure should incorporate three elements: the inherent *attractiveness* of services; their *proximity* to the population of interest; and the effect of competing populations. The traditional method of treating such concepts is to develop a measure of the *accessibility* of the ward to NHS services. This is achieved here using the ideas of spatial interaction modelling described by Wilson (1974), but substituting the notion of a *hospital episode* for the conventional spatial interaction phenomenon of a *trip*.

The standard spatial interaction model is of the form:

$$T_{id}=gP_iS_d\,f(c_{id}) \tag{1}$$

where T_{id} is the number of interactions (hospital episodes per year) between residential zone i and destination d;
P_i is some measure of the effective population of zone i;
S_d is some measure of the size or attractiveness of destination d;
c_{id} is some measure of distance (or time) between i and d;
f(.) is a distance decay or deterrence function;
g is a gravitational constant.

Then the total number of interactions (hospital episodes) T_i generated by zone i per year is given by

$$T_i=gP_i\sum_d S_d\,f(c_{id}) \tag{2}$$

and the number of episodes T_d attracted to destination (hospital) d is

$$T_d=gS_d\sum_i P_i\,f(c_{id}) \tag{3}$$

Now in this study each hospital (destination) is limited in the number of patients it can

treat. That is, the model is "attraction constrained" (Batty, 1976, p39). It is therefore necessary to introduce a balancing factor B_d into the model for each destination d, so that (1) is rewritten

$$T_{id}=gP_iB_dS_d\,f(c_{id}) \tag{4}$$

where

$$B_d=\frac{1}{\sum_i P_i\,f(c_{id})} \tag{5}$$

Introduction of the factor B_d ensures that the influence of competing populations is properly modelled.

Then the accessibility A_i of zone i to hospital facilities can be given by the ratio of predicted number of episodes in relation to population, which is represented by the expression

$$A_i=\left(\sum_d T_{id}\right)/P_i=\sum_d B_dS_d\,f(c_{id})=\sum_d\left(\frac{S_d\,f(c_{id})}{\sum_r P_r\,f(c_{rd})}\right) \tag{6}$$

Expression (6) models the *relative* accessibility of residents in zone j to all hospital resources, given the availability of beds (S_d), the distance to the hospitals (c_{id}) and the competition from local populations. It is a *distance weighted* form of the simple ratio "beds per head".

Thus in order to calculate the accessibility of residential zones, it is first necessary for each hospital to calculate the index B_d . Once the form of the deterrence function has been chosen, this is straightforward. Choice of measures for P_i and S_d is also straightforward: population and beds serve as reasonable proxies for demand (people) and supply (episodes). (Note that *demographic* determinants of utilization were treated

elsewhere in this study, so that the population did not have to be weighted by need. The measure A_i is merely intended to give a measure of relative inpatient provision.) The measure of distance c_{id} should ideally be a measure of *perceived* distance, or possibly journey time. However, in this study the only available distance measures were straight line (or crow fly) distances, so these had to be used. A standard intrazonal cost was added to each distance.

Finally, possibly the most troublesome aspect of modelling is the choice of deterrence function f(.). Scrutiny of the spatial location literature suggests a wide range of possible functional forms. Haggett, Cliff and Frey (1977) describe two in widespread use:

$$f(c)=e^{-\beta c^{\alpha}} \qquad f(c)=c^{-\beta} \tag{7}$$

where c is distance and α and β are parameters to be estimated.

The distance function can be calibrated using a gravity model of the sort described by Batty (1976), in which case the parameters α and β are chosen to maximize a suitable likelihood function. Because we had no information about hospital of treatment we could not calibrate a gravity model. The original Newtonian model of physical gravitation uses the second of the functional forms with $\beta = 2$ (the inverse square law). Unfortunately, in modelling social phenomena, there is no guarantee that such a neat result exists. As a result, it was necessary to appeal to previous studies and judgement to model deterrence.

Batty uses both functional forms for subregional modelling. Using the first, he sets $\alpha = 1$, and estimates β by an iterative process such that modelled mean trip length equals observed mean trip length. Values of between 0.1 and 0.3 are found. Using the second, values of β between 1.5 and 2.5 are found. In another study, Foot (1981) uses the first functional form with $\alpha = 1$ and $\beta = 0.2$. The relevance of these values to the current study is limited because of the very particular type of spatial interaction being modelled. Indeed, we might expect that - for different types of NHS referral - different types of deterrence might occur. The only directly relevant work is the study of London hospitals

reported by Mayhew (1986). However, he gives no values for the deterrence function parameters. In general, relatively minor conditions might be expected to exhibit high elasticity with respect to distance (high values of β) while lower values of β might obtain for, say, regional specialties. Bearing in mind that we wished to arrive at a relatively broad brush measure of accessibility, it was unnecessary to model such subtleties.

Instead, accessibility was modelled using the following two deterrence functions:

$$f(d)=e^{-0.2c} \qquad f(d)=c^{-2} \tag{8}$$

The measures of accessibility implicit in these choices were examined to check that they were reasonable. It was eventually decided to use an inverse square deterrence function with intrazonal cost of 10 kilometres.

References

Batty, M. (1976), *Urban modelling: algorithms, calibrations, predictions*, Cambridge: Cambridge University Press.

Foot, D., (1981) *Operational urban models*, London: Methuen.

Haggett, P., Cliff, A. D. and Frey, A. (1977), *Locational models*, London: Edward Arnold.

Mayhew, L. (1986), *Urban hospital location*, London: Allen and Unwin.

Wilson, A. G. (1974) *Urban and regional models in geography and planning*, Chichester: Wiley.

APPENDIX D: THE STATISTICAL MODELLING STRATEGY

The central hypothesis proposed in this study has been that health care utilisation depends on 'needs' and health care supply. Health care supply in turn depends on needs and utilisation. Therefore a simultaneous equations model is suggested. The equation of interest is the utilisation equation in which Two Stage Least Squares (2SLS) estimation is employed. The modelling methodology used follows three main stages.

The first stage involves testing the assumption that supply is in fact endogenous in the utilisation equation. This procedure is covered in Section D1. Providing the assumption of endogeneity of the supply variables is confirmed by the first stage the next stage involves estimating the equation of interest by 2SLS. As with Ordinary Least Squares (OLS) it is important to test for misspecification of the equation estimated. The procedure for testing for misspecification in this study is outlined in Section D2. The approach adopted in this study followed Hendry's 'general to specific' schema (Hendry 1987). A general model is estimated which contains the vector of endogenous regressors and many predetermined regressors. This model is tested for misspecification, and then a series linear restrictions are tested on the coefficients of the predetermined regressors to attempt to find a more parsimonious model. Each restriction is tested for its statistical validity, and the smaller model is also tested for misspecification. The test for linear restrictions in a 2SLS model is described in Section D3. The third stage of the modelling process involves diagnostic checking of the models for endogeneity in the instrument set (Section D4), and testing for heteroscedasticity (Section D5).

D1. Testing for endogeneity

If we wish to estimate the first equation of the structural model which takes the following form

$$U=f(H,N',S)$$

$$S_1=f(H,N,U)$$

$$\vdots \qquad \vdots$$

$$S_p=f(H,N,U).$$

Where H is a vector of 'health needs'

N is a vector of 'social needs'

U is a measure of utilisation

S is a vector of supply measures (p measures available)

N' is a subset of the social needs vector N.

We need to test the hypothesis that the supply measures are actually endogenous in the first equation.

1. Run OLS regressions for the p supply equations

$$S_1 = f(H,N) + e_1$$
$$\vdots \qquad \vdots$$
$$S_p = f(H,N) + e_p$$

where $e_1,\ldots,e_p$ are the residuals from these regressions.

2. Take the first equation of the model

$$U = f(H,N',S)$$

and add the residuals from the OLS regressions in step 1 as regressors on the right hand side of the equation. Next run an OLS regression on this new equation which includes these residuals. The estimated equation will look like this

$$U = f(H,N',S,e_1,\ldots,e_p) + u$$

where u are the residuals from this new regression.

3. Test the hypothesis of the supply variables being endogenous in the (p+1) structural equation model by using an F test. Therefore all supply variables are

tested together. Under the null hypothesis (H_0) the supply variables are exogenous, therefore the equation of interest under H_0 is given by

$U=f(H,N',S)$ This is the **restricted** equation (R)

Under the alternative hypothesis (H_1) the supply variables are endogenous and the equation is given by

$U=f(H,N',S,e_1,\ldots,e_p)$ This is the **unrestricted** equation (UR)

The test statistic is the of the usual form for an F test

(i) Calculate the residual sum of squares from the estimation of the restricted equation (RSS_R)

(ii) Calculate the residual sum of squares from the estimation of the unrestricted equation (RSS_{UR})

The test statistic is given by

$$\frac{(RSS_R - RSS_{UR})/(no.\ of\ restrictions)}{(RSS_{UR})/(no.\ observations - no.\ regressors\ unrestricted\ equation)} \sim F_{(p,n-m1)}$$

Where n is the number of observations,
p is the number of regressors tested for endogeneity,
m1 is the number of regressors in the unrestricted equation.

4. *If the test statistic is greater than the critical value then H_0 is rejected. That is the supply variables are endogenous.*

D2. Testing for misspecification under 2SLS

The test proposed is a **general** test for misspecification of a single equation of interest in a simultaneous equation model (Godfrey 1988). This means the test consists of a null hypothesis that the equation of interest is correctly specified, and a general alternative hypothesis. The test therefore indicates whether the null hypothesis (the equation is correctly specified) should be rejected or not.

1. Save the residuals generated from the estimation of the utilisation equation (the equation of interest) under 2SLS. Let these residuals be called e_{2SLS}. These are the residuals from the regression

$$U=f(H,N',S) + e_{2SLS}$$

2. Regress the residuals e_{2SLS}, using OLS, against all the full instrument set used in the 2SLS regression. This is known as the auxiliary regression, and is given by:

$$e_{2SLS}=f(W) + v$$

where W is a vector of the instruments used in the 2SLS regression, and v is the error term in the auxiliary regression. W contains all the needs variables in the model, H and N, which are predetermined variables.

3. Derive the explained sum of squares from the auxiliary regression run under step 2. Let this be known as ESS_A.

4. Derive the residual sum of squares from the original 2SLS regression used for step 1. Let this be known as RSS_{2SLS}.

$$RSS_{2SLS}= \Sigma(e_{2SLS}^2)$$

Divide RSS_{2SLS} by:

(the number of wards) - (the number of regressors in the utilisation equation).

The value derived is the estimated error variance for the utilisation equation in the 2SLS estimation. This is written $\hat{\sigma}^2$, which is given by:

$$\hat{\sigma}^2 = \frac{\Sigma(e^2_{2SLS})}{(no.\ of\ wards\ -\ no.\ of\ regressors\ utilsation\ equation)}$$

5. The test statistic is:

$$\frac{ESS_A}{\hat{\sigma}^2}$$

which was found in steps 3 and 4.

6. The test statistic is asymptotically distributed chi squared, with the degrees of freedom given by:

(Number of instruments) - (Number of regressors in the utilisation equation)

$$\frac{ESS_A}{\hat{\sigma}^2} \rightarrow \chi^2_{df}$$

7. *If the test statistic is greater than the critical value found in statistical tables then the model is misspecified.*

This may indicate misspecification of the equation of interest: that is, omitted variables or incorrect functional form. It may also indicate that some of the predetermined variables used in the equation may in fact be endogenous.

D3. Testing for linear restrictions

The methodology employed is 'general to specific' in developing the utilisation equation for the simultaneous equation model. This involves determining a general utilisation equation at the outset which includes large numbers of regressors, and this is referred to as the **unrestricted** model in the following discussion. This equation is tested for misspecification as under Section D2 above.

The aim is to find a 'smaller' equation including less regressors to explain variations in utilisation rates across our set of observations. This is achieved by imposing sets of linear restrictions to the general model, that is setting some of the regressors coefficients to zero. The new, smaller model is referred to as the **restricted** model. Once a smaller model is decided upon it must first be tested for misspecification as in Section D2 above. Providing both unrestricted and restricted models are specified correctly then one can test to see if these restrictions are valid, and if these restrictions provide a model which more accurately explains variation in utilisation rates.

1. Run the misspecification test outlined above in Section D2 on both the unrestricted (general) model and on the restricted (smaller) model.

2. Derive ESS_A, the explained sum of squares from the auxiliary regression in Section D2 step 3, for the unrestricted model. Let this be known as $ESS_{A(UR)}$.

3. Derive ESS_A, the explained sum of squares from the auxiliary regression in Section D2 step 3, for the restricted model. Let this be known as $ESS_{A(R)}$. The instrument set, W, is kept the same throughout the testing procedure.

4. Derive the estimated error variance for the utilisation equation for the restricted model. This easily found under calculation made in Section D2 step 4. This is written $\hat{\sigma}_R^2$.

5. The test statistic is given by the following formula:

$$\frac{ESS_{A(R)} - ESS_{A(UR)}}{\hat{\sigma}_R^2}$$

which is 'similar' to the usual F test used for testing linear restrictions.

6. The test statistic is asymptotically distributed chi squared, and the degrees of freedom are the number of restrictions placed on the general model.

$$\frac{ESS_{A(R)} - ESS_{A(UR)}}{\hat{\sigma}_R^2} \rightarrow \chi^2_{df}$$

7. *If the test statistic is greater than the critical value given by statistical tables then the restricted model is rejected.*

D4. Testing for endogenous instruments

Some of the instrument set W may actually be endogenous in the equation of interest. Therefore it is prudent to test any variables in the instrument set that we think could be endogenous. To test for endogeneity of these variables it is important to separate them out from the supply variables which were tested for endogeneity in Section D1. Testing any instruments with the supply variables could indicate such instruments are endogenous when in fact they are not.

Assume for example that the instruments suspected of being endogenous are SMR and SIR.

1. If the model is:

$$U=f(H,N',S)$$

$$S_1=f(H,N,U)$$

$$\vdots \qquad \vdots$$

$$S_n=f(H,N,U)$$

where H and N form the set of instruments W, one may suspect that some of the N' (which is a subset of N) variables in the utilisation equation are in fact endogenous. Therefore the utilisation equation could be rewritten:

$$U=f(H,N'',Y,S)$$

where N'' are the definitely predetermined variables in N', and Y are the variables contained in N' that we suspect to be endogenous.

To carry out the test for endogeneity on Y one must have a set of instruments W^* which is valid **whether Y is endogenous or predetermined**. W^* is a subset of the original instrument set W, since the instrument set cannot contain variables that are suspected to be endogenous. One also must still have more instruments than parameters to be estimated in the utilisation equation.

2. Under the null hypothesis (H_0) Y is predetermined. Therefore under the null

$$W=f(W^*,Y)$$

is a valid instrument set, as Y is predetermined and can therefore be used as an instrument. Regress Y on W^* using OLS and save the residuals, where the

residuals are given by e_A. The regression equation is:

$$Y=f(W^*) + e_A$$

This is called the auxiliary regression.

3. The procedure is to test the **joint** significance of e_A in the original utilisation equation under the null hypothesis that Y is predetermined. Therefore estimate the equation below using 2SLS and the original set of instruments W:

$$U=f(H,N'',Y,S,e_A) + e_1$$

So the residuals from the auxiliary regression are acting as independent variables in this new regression.

Save the residuals e_1

4. Run the regression (which is the **unrestricted** model)

$$e_1=W + e_1^*$$

using OLS and making sure the full instrument set W is used. Calculate the explained sum of squares for this regression: ESS_{UR}. Derive the explained sum of squares for this regression (ESS_{UR})

5. Estimate the equation

$$U=f(H,N'',Y,S) + e_2$$

using 2SLS again, and the same instrument set W. Note this is just the original

utilisation equation to be estimated

$$U=f(H,N',S)$$

as N'' and Y go to make up the vector N'. Save the residuals e_2

6. Run the regression (which is the **restricted** model)

$$e_2=W + e_2^*$$

using OLS and making and the full instrument set W. Calculate the explained sum of squares for this regression: ESS_R

7. Derive the estimated error variance for the restricted model. This is given by

$\hat{\sigma}_R^2$, which as before is found from the calculation:

$$\hat{\sigma}^2 = \frac{\Sigma(e_2^{*2})}{(no.\ of\ wards\ -\ no.\ of\ regressors\ utilsation\ equation)}$$

noting that the residual sum of squares $\Sigma(e_2^{*2})$ refers to the **restricted** equation.

8. The test for endogeneity uses the "F test" in Section D3 above for testing linear restrictions imposed on a 2SLS equation. Therefore the test is of the joint significance of e_1 in a 2SLS regression. The test statistic is:

$$\frac{ESS_{A(R)} - ESS_{A(UR)}}{\hat{\sigma}_R^2}$$

9. This test statistic is asymptotically distributed chi squared with the degrees of

freedom determined by the number of restrictions imposed (the number of variables being tested for endogeneity).

$$\frac{ESS_{A(R)} - ESS_{A(UR)}}{\hat{\sigma}^2_R} \rightarrow \chi^2_{df}$$

D5. Testing for heteroscedasticity

1. Assume no heteroscedasticity exits in the other structural equations in the model. This makes the use of tests for heteroscedasticity under OLS possible for our equation estimated under 2SLS.

2. Run a 2SLS regression on the utilisation equation and save the residuals generated. The regression will therefore be of the form:

$$U = f(H,N',S) + e_{2SLS}$$

3. Square these residuals to get e^2_{2SLS}

4. Run an unrestricted auxiliary regression using e^2_{2SLS} as the dependent variable. The independent variables will be those variables in the model which are suspected of causing heteroscedasticity in the error variance. The unrestricted auxiliary regression should include an intercept term. An example may be to test the regional dummy variables, therefore the test would be as follows: run the OLS regression

$$e^2_{2SLS} = \alpha_0 + \alpha_1 x_1 + ... + \alpha_{14} x_{14} + v$$

where the intercept term is α_0, and the 14 dummy variables for the regions are

$x_1,\ldots,x_{14}$. The error term v on the unrestricted auxiliary regression is assumed to obey all the classical linear regression assumptions, and is therefore assumed to be homoscedastic.

5. A large sample analogue of the F test can be used to test whether the variables thought to cause the heteroscedasticity in the above equation are jointly significant. The test is based on the Lagrange Multiplier type test. The test statistic is:

$$nR^2 \sim \chi^2_k$$

where n is the number of wards (observations), and k is the number of regressors in the equation in step 4 (in the above example k=14).

6. *If the test statistic is greater than the critical value then H_0 is rejected. That is the errors in the 2SLS utilisation equation are heteroscedastic.*

References

Hendry D F (1987) *Econometric methodology: a personal perspective* in Bearly T F (ed) *Advances in Econometrics*, Cambridge: Cambridge University Press

Godfrey LG (1988) *Misspecification tests in Econometrics*, Cambridge: Cambridge University Press.

APPENDIX E: MULTI-LEVEL ESTIMATION

Districts and regions may affect utilization. Wards in the same district or region may tend to be more similar than wards in different districts or regions. This leads to clustering at the regional and ward level. OLS and 2SLS ignore this clustering. This can lead to biased estimates of parameters and standard errors. Multilevel modelling takes account of district and regional effects. This appendix briefly describes the differences between multilevel modelling and OLS regression.

Let us take a simple example where we are examining the relationship between utilisation (U) and SMR (S). The OLS model would be

$$U_i = \alpha + \beta S_i + e_i \qquad (1)$$

No account is taken of districts and regions. A multilevel model allows the intercepts and if desired the slopes to vary across regions and districts. We can write a multilevel model which allows the intercept to vary across districts as:

$$U_{ij} = \alpha_j + \beta S_{ij} + e_{ij} \qquad (2)$$

where U_{ij} is the utilisation of the i'th ward in the j'th district, α_j is the intercept for the j'th district, S_{ij} is the SMR of the i'th ward in the j'th district and e_{ij} is a ward level random variable, which is commonly thought of as the error or residual term.

OLS models have only one random or residual term. What distinguishes multilevel models from OLS models is that they may have more than one random term. In equation (2) we write

$$\alpha_j = \alpha + v_j \qquad (3)$$

This expresses the intercept for the j'th district as an average value (α) plus a random departure (v_j). We can therefore rewrite equation (2) as

$$U_{ij} = \alpha + \beta S_{ij} + v_j + e_{ij} \qquad (4)$$

The model now has two fixed parameters (α and β) and two random parameters (v_j and e_{ij}). We can estimate *var* (v_j) = σ_v^2 and *var* (e_{ij}) = σ_e^2, which are the between district and between ward variances. The ratio of the between district variance to the total variance.

$$\frac{\sigma_v^2}{\sigma_v^2 + \sigma_e^2} \qquad (5)$$

is the intra-district correlation. The higher this correlation the stronger the clustering. Values of over 0.2 indicate strong clustering. In the data analyzed in this report the intra-district correlation was of the order of 0.3.

The district and ward level random variables are assumed to be normally distributed.

It is possible using OLS analysis of covariance techniques to estimate direct effects by fitting dummy variables for each district. This has two draw backs. Firstly, we have 186 districts so we have to fit a large number of dummy variables. Second, the intercept for the j'th district is estimated from the data for only that district. Multilevel models estimate a distribution of district effects so data from all the other 185 districts inform the estimate for any particular district.

We could gain an estimate of σ_v^2 by taking the variance of the 186 dummy coefficients. However this estimate would be biased, inefficient and non maximum likelihood. Multilevel model estimate is an iterative procedure where the estimates of the random parameters at iteration t-1 are used to weight the computation of the estimates of fixed and random parameters at iteration t. This weighted iterative procedure leads to unbiased, efficient maximum likelihood estimates.

Multilevel models are easily extended. We can develop equation (2) so that both the intercepts and slopes are allowed to vary across districts. This model is

$$U_{ij} = \alpha_j + \beta_j S_{ijk} + e_{ij} \quad (6)$$

σ and β are now modelled as a fixed value plus a random departure at the district level. This model can be developed by adding more explanatory variables whose coefficients may also vary across districts.

APPENDIX F: OPTIONS FOR FURTHER WORK

The work described in this report was limited by a number of factors relating to the quality, detail and level of aggregation of the data. These factors are considered in this Appendix, which was prepared after consultation with Professor Brian Jarman of St Mary's Hospital Medical School, Gwyn Bevan of London Economics and Ken Judge and Nick Mays of the King's Fund Institute.

(i) *Data Quality*

There are acknowledged problems with the data that are collected in the Hospital Episode system. Some of these problems can be solved. For example, HES should be able to provide reliable information on utilization on long unfinished episodes. If it were possible to use a census of all bed use in a year, then models could be developed with more confidence in the non-acute sector. It should be possible to include more data on the 'needs' profile of patients admitted (other than age or sex). In addition, if the HES data provided were to be extended to include hospital of treatment, it would be possible to calculate utilization rates for different hospitals by the population of each ward, thereby permitting the use and calibration of spatial interaction models. This would facilitate the development and testing of alternative accessibility indices - allowing, for example, for differential decay between the local district general hospital and other hospitals - and would assist in the analysis of catchment areas.

(ii) *Effect of Time*

Whilst the theoretical model posited a complex lag structure to the relationships between demand, supply and utilization, these could not be explored with cross-sectional data from one year. If it were felt to be worth persisting with the small area approach, one possible extension would be to add a time dimension, with HES and population data for a sequential series of time periods. Extensions to the model of this sort would, however, be constrained by the fact that data on most variables are only available nationally from the decennial census.

(iii) *Level of Aggregation*

The results of the current exercise - and especially of the multilevel modelling - have suggested that the models may have been more robust if data relating to the individual as well as the ward and district were incorporated into the analysis. Because wards are not homogenous, it is unlikely that their social characteristics necessarily represent the characteristics of the individuals using the hospital. Even a move down to enumeration district level would be of limited value because of the Barnardization of Census data (the practice of randomly adding -1, 0 or +1 to counts to protect confidentiality). In our view, a combination of individual data and associated area characteristics is likely to be more powerful than any small area analysis in explaining variations in utilization.

(iv) *Utilization as an index of need*

This entire study was predicated on the assumption that utilization of NHS inpatient resources is a good predictor of health care need. For many reasons, this assumption may be suspect. Some groups of the population may be systematically excluded from NHS services, while others may "capture" more NHS resources than their clinical need justifies. There is a clear need for research to establish whether utilization is a legitimate predictor of need.

Recent research suggests that for many types of care there is a great deal of variation in utilization which cannot be explained by variations in morbidity. Where this is the case, there are therefore likely to be high levels of inappropriate care. In these circumstances it becomes difficult to disentangle differences in utilization due to health care needs, and differences due to other extraneous factors. This suggests a need for research which examines for particular types of care how variations in utilization are related to variations in clinical and social need, and which analyzes "unexplained" variations in utilization. This research would cast light on the extent to which, for specific conditions, the use of utilization as an index of need is justified.

(v) *Use of resources*

Currently, there is little information on the efficacy with which resources are actually used. As currently organised, data on hospital episodes cannot provide information on the process quality of care or on the outcomes of that care. Given the complexity of measuring both process quality or outcomes, there are no easy solutions: but without that information, it is impossible for a district purchaser to assess the appropriateness of the care packages being provided from available resources, or indeed the equity of the distribution within the district. These issues cannot be dealt with by a statistically derived average formula but should be major considerations for purchasers.

(vi) *Link with other sources of care*

This work was designed to develop a formula for the allocation of HCHS resources. As DHAs and FHSAs merge to form commissioning authorities, there is a need to consider HCHS and GMS allocations together, possibly leading to a common system of resource allocation (see Chapter 3). Moreover, as this analysis has indicated, the use of inpatient care is associated with the provision of primary care, private health care and nursing and residential homes. And, given the importance of socio-economic variables in the model, it is reasonable to assume that expenditure by local authorities will also have an impact on the need for hospital services. The ways in which these and other social expenditures should be taken into account when considering the allocation of resources for health care requires more thought and research.

While the above issues can to some extent be addressed by enhancing existing data sources and further research effort, some of the issues raised suggest the need for different types of data to those currently available. In particular, information is required on:

(a) what are the needs for health care and how are these related to individual characteristics and, eventually, locality circumstances?

(b) what is the response of the health care system to the health care needs of

the individual (whether or not expressed) and how does that response vary according to area and physician characteristics?

(c) what are the outcomes for the individual?

In principle, a complete Management Information System (MIS) covering all transactions within the health care system could provide some data on (b) and (c) but not on (a).

In summary, the most important need is to understand who (in terms of their socio-economic and other characteristics) experiences which kinds of illness (in terms of their morbidity profile), what treatments they receive, and what levels of care are appropriate. The only way to make this principle operational would be to carry out a panel study of a large number of households. This would examine the extent to which different sorts of people (in terms of health and social circumstances) with different levels of local provision of health and social services could benefit from health care intervention. Using the results of such research, it would then be possible to build up profiles of the needs of areas by inferring the needs of individuals within those areas. This issue becomes particularly important with the shift towards community-based provision, such as GP fundholding, for which budgets have to be set for very small populations. The needs indicators for these small populations can have very much higher variation than those for health authorities, and so the importance of an accurate and sensitive basis for setting budgets becomes even greater. Without such a cohort study the principle of delegating budgets in an equitable fashion to small areas below District level may be fatally compromised.